PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association

Seattle, WA; 1-4 November 2018

Version: 30 January 2019

PhilSci ARCHIVE

Table of Contents

Mixed-Effects Modeling and Non-Reductive Explanation

(4975 words)

Abstract: This essay considers a mixed-effects modeling practice and its implications for the philosophical debate surrounding reductive explanation. Mixed-effects modeling is a species of the multilevel modeling practice, where a single model incorporates simultaneously two (or even more) levels of explanatory variables to explain a phenomenon of interest. I argue that this practice makes the position of explanatory reductionism held by many philosophers untenable, because it violates two central tenets of explanatory reductionism: single level preference and lower-level obsession.

1. Introduction

Explanatory reductionism is the position which holds that, given a relatively higher-level phenomenon (or state, event, process, etc.), it can be reductively explained by a relatively lower-level feature (Kaiser 2015, 97; see also Sarkar 1998; Weber 2005; Rosenberg 2006; Waters 2008).[1] Though philosophers tend to have slightly different conceptions of the position, two central tenets of the position can still be extracted:[2]

> **Single level preference:** a phenomenon of interest can be fully explained by invoking features that reside at a single, well-defined level of analysis (e.g., molecular level in biology).

[1] According to Sarkar (1998), explanatory reduction is an epistemological thesis which is distinguished from constitutive (ontological) and theory reductionism theses. Kaiser further distinguishes two sub-types of explanatory reduction: (a) "a relation between a higher-level explanation and a lower-level explanation of the same phenomenon" (2015, 97); (b) individual explanations, i.e., given a relatively higher-level phenomenon, it can be reductively explained by a relatively lower-level feature (*Ibid*., 97). This essay will focus on the second sub-type. Besides, when referring to levels I mean either hierarchical organization such as universities, faculties, departments etc., or functional organization such as organs, tissues, cells etc. When referring to scales I mean spatial or temporal scaling where levels are not so clearly delimited.

[2] Similar summary of the position can be found in Sober (1999).

> **Lower-level obsession:** lower-level features always provide the most significant and detailed explanation of the phenomenon in question, so a lower-level explanation is always better than a higher-level explanation.

Philosophers sometimes express these two tenets explicitly in their work. For example, Alex Rosenberg holds that "[…] there is a full and complete explanation of every biological fact, state, event, process, trend, or generalization, and that this explanation will cite only the interaction of macromolecules to provide this explanation" (Rosenberg 2006, 12). Marcel Weber expresses a similar idea in his explanatory hegemony thesis, according to which it's always some lower-level physicochemical laws (or principles) that ultimately do the explanatory work in experimental biology (Weber 2005, 18-50). John Bickle attempts to motivate a 'ruthless' reduction of psychological phenomena (e.g., memory) to the molecular level (Bickle 2003).

However, many philosophers have questioned the plausibility of the position on the basis of scientific practice (Hull 1972; Craver 2007; Bechtel 2010; Brigandt 2010; Hüttemann and Love 2011; Kaiser 2015). To counter that position, some authors have pointed to the relevance of an important practice that has not received sufficient attention before: multiscale or multilevel modeling or sometimes called integrative modeling approach, where a set of distinct models ranging over multiple levels or scales—including the macro-phenomenon level/scale—are involved in explaining a (often complex) phenomenon of interest

(Mitchell 2003, 2009; Craver 2007; Brigandt 2010, 2013a, 2013b; Knuuttila 2011; Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017). Often these models work together by providing diverse constraints on the potential space of representation (Knuuttila and Loettgers 2010; Knuuttila 2011; Green 2013).

This multilevel modeling surely casts some doubt on explanatory reductionism, for it seems unclear what reductively explains what—all those facts in the set of models ranging over different levels/scales are involved in doing some explanatory work. However, there is a species of multilevel modeling that has slipped away from most philosophers' sights: mixed-effects modeling (MEM hereafter)—also called multilevel regression modeling, hierarchical linear modeling, etc.—in which a single model incorporating simultaneously two (or even more) levels of variables is used to explain a phenomenon. For a mixed-effects model to explain, features of the so-called reducing and reduced levels must be simultaneously incorporated into the model, that is, they must go hand in hand.

MEM deserves special attention because it sheds new light on the reductionism-antireductionism debate by showing that (a) a mixed-effects model violating the two central tenets of explanatory reductionism can provide successful explanation, and (b) a single mixed-effects model without integrating with other epistemic means can also provide such successful explanation. Therefore, MEM first further challenges the explanatory reductionist position, and

second offers a novel perspective bolstering the multilevel/multiscale integrative approach discussed by many philosophers.

The essay proceeds as follows. Section 2 discusses the challenges faced by the traditional single-level modeling approach, and examines the reasons why the MEM approach is preferable in dealing with these challenges. Section 3 describes a MEM practice using a concrete model. Section 4 elaborates on the implications of MEM for the explanatory reductionism debate. Finally, Section 5 considers potential objections to my viewpoint.

2. Challenges to Reductive Explanatory Strategies

In many fields (e.g., biological, social and behavioral sciences) scientists find that the data collected show an intrinsically hierarchical or nested feature. Consider a simple example: we might be interested in examining relationships between students' achievement at school (A hereafter) and the time they invest in studying (T).[3] In conducting such a research, we might collect data from different classes (say 5 classes in total), with each class providing the same number of samples (say 10 students in each class). The data collected among classes might be taken for granted to be independent. Then we may use certain traditional statistical techniques such as ordinary least-squares (OLS) to analyze the data and build a linear relationship between A and T.

[3] For scientific studies of this kind, see Schagen (1990), Wang and Hsieh (2012), and Maxwell et al. (2017).

However, this single-level reductive analysis can lead to misleading results, because it ignores the possibility that students within a class may be more similar to each other in important aspects than students from different classes. In other words, each group (class) may have its own features relevant to the relationship between A and T that the other groups lack. Hence, the data collected from the students are in fact not independent, i.e., the subjects are not randomly sampled, because the individuals (students) are clustered within groups (classes). In technical terms, we say our analysis may fall prey to the *atomistic fallacy* where we base our analysis solely on the individual level—i.e., we reduce all the group-level features to the individuals. Therefore, traditional OLS techniques such as multiple regression cannot be employed in this context, because the case under consideration violates a fundamental assumption of these techniques: the independence of observations (Nezlek 2008, 843).

Conversely, we may face the same problem the other way around if we fail to consider the inherently nested nature of the data. Consider the student-achievement-at-school case again. We may observe that in classes where the time of study invested by students is very high, the achievements of the students are also very high. Given such an observation, we may reason that students who invest a lot of time in studying would be more likely to get higher achievements at school. However, this inference commits the *ecological fallacy*, because it attributes the relationship observed at the group-level to the individual-level (Freedman 1999). The individuals may exhibit within-group differences that the single group-level analysis fails to capture. In technical terms, this inference flaws

because it reduces the variability in achievement at the individual-level to a group-level variable, and the subsequent analysis is solely based on group's mean achievement results (Heck and Thomas 2015, 3). Again, traditional statistical techniques such as multiple regression cannot be employed in this context.

In sum, a single-level modeling approach that disrespects the multilevel data structure can commit either an atomistic or an ecological fallacy. Confronted with these problems, one response is to 'tailor' the traditional statistical techniques by, e.g., adding an effect variable to the model which indicates the grouping of the individuals. However, many have argued that this approach is unpromising because it may give rise to enormous new problems (Luke 2004; Nezlek 2008; Heck and Thomas 2015). Alternatively, scientists have developed a new framework that takes the multilevel data structure into full consideration, i.e., the MEM approach, to which we now turn.

3. Case Study: A Mixed-Effects Model

Depending on different conceptual and methodological roots we have two broad categories of MEM approaches: the multilevel regression approach and the structural equation modeling approach. The former usually focuses on direct effects of predictor variables on (typically) a single dependent variable, while the latter usually involves latent variables defined by observed indicators (for details see Heck and Thomas 2015). For the purpose of this essay's arguments, I will concentrate on the first kind.

Consider the student-achievement-at-school example again. Since students are typically clustered in different classes, a student's achievement at school may be both influenced by her own features (e.g., time invested in studying) and her class's features (e.g., size of the class). Hence here comes two levels of analysis: the individual-level (level-1) and the group-level (level-2), and individuals ($i=1,2,\ldots,N$) are clustered in level-2 groups ($j=1,2,\ldots,n$).[4] Now suppose that students' achievements at school are represented as scores they get in the exam. The effect of time invested in studying on scores can be described as follows:

$$Y_{ij}=\beta_{0j}+\beta_{1j}X_{ij}+\varepsilon_{ij} \qquad (1)$$

where Y_{ij} refers to the score of individual i in the jth group, β_{0j} is a level-1 intercept representing the mean of scores for the jth group, β_{1j} a level-1 slope (i.e., different effects of study time on scores) for the predictor variable X_{ij}, and the residual component (i.e., an error term) ε_{ij} the deviation of individual i's score from the level-2 mean in the jth group. Equation (1) looks like a multiple regression model; however, the subscript j reveals that there is a group-level incorporated in the model. It can also be seen from this equation that both the intercept β_{0j} and slope β_{1j} can vary across the level-2 units, that is, different groups can have different intercepts and slopes.

[4] Note that, for instructive purposes, our case involves only two levels; however, the MEM approach can in principle be extended to many more levels.

The most remarkable thing of MEM is that we treat both the intercept and slope at level-1 as dependent variables (i.e., outcomes) of level-2 predictor variables. So here we write the following equations expressing the relationships between the level-1 parameters and level-2 predictors:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad (2)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \quad (3)$$

where β_{0j} refers to the level-1 intercept in level-2 unit j, γ_{00} denotes the mean value of the level-1 intercept, controlling for the level-2 predictor W_j, γ_{01} the slope for the level-2 variable W_j, and u_{0j} the error (i.e., the random variability) for unit j. Also, β_{1j} refers to the level-1 slope in level-2 unit j, γ_{10} the mean value of the level-1 slope controlling for the level-2 predictor W_j, γ_{11} the effect of the level-2 predictor W_j, and u_{1j} the error for unit j.

Equations (2) and (3) have specific meanings and purposes. They express how the level-1 parameters, i.e., intercept or slope, are functions of level-2 predictors and variability. They aim to explain variations in the randomly varying intercepts or slopes by adding one (or more) group-level predictor to the model. These expressions are based on the idea that the group-level characteristics such as group size may impact the strength of the within-group effect of study time on

scores. This kind of effect is called a *cross-level interaction* for it involves the impact of variables at one level of a data hierarchy on relationships at another level. We will discuss this in detail in the next section.

Now we combine equations (1), (2) and (3) by substituting the level-2 parts of the model into the level-1 equation. We finally obtain the following equation:

$$Y_{ij}=\left[\gamma_{00}+\gamma_{10}X_{ij}+\gamma_{01}W_j+\gamma_{11}X_{ij}W_j\right]+\left[u_{1j}X_{ij}+u_{0j}+\varepsilon_{ij}\right] \qquad (4)$$

This equation can be simply understood that Y_{ij} is made up of two components: the fixed-effect part expressed by the first four terms and the random-effect part expressed by the last three terms. Note that the term $\gamma_{11}X_{ij}W_j$ denotes a cross-level interaction between level-1 and level-2 variables, which is defined as the impact of a level-2 variable on the relationship between a level-1 predictor and the outcome Y_{ij}. We have 7 parameters to estimate in (4), they are four fixed effects: intercept, within-group predictor, between-group predictor and cross-level interaction, two random effects: the randomly varying intercept and slope, and a level-1 residual.

Now a mixed-effects model has been built, and the next step is to estimate the parameters of the model. However, we will skip this step and turn to explore the philosophical implications of the modeling practice relevant to the explanatory reductionism debate.

4. Implications for the Explanatory Reductionism Debate

Looking closely into the MEM practice, we find that a couple of important philosophical implications for the explanatory reductionism debate can be drawn.

4.1. All levels are indispensable

The first, and most obvious, feature of MEM is that it routinely involves many levels of analysis in a single model, and all these levels are indispensable to the model in the sense that no level can be reduced to or replaced by the other levels. These levels consist of both the so-called reducing level in the reductionist's terminology, typically a lower-level that attempts to reduce another level, and the reduced level, typically a higher-level to be reduced by the reducing level. In our student-achievement-at-school case, for example, a reductionist may state that the group-level will be regarded as the reduced level whereas the student-level as the reducing level.

The indispensability of each level in the model can be understood in two related ways. First, due to the nested nature of data, only when we incorporate different levels of analyses to the model can we avoid either the atomistic or ecological fallacy discussed in Section 2. As discussed in the student-achievement-at-school example where students are clustered in different classes (in the manner that students from the same class may be more similar to each other in important aspects than students from different classes), reducing all the analyses to the level of individual students can simply miss the important

information associated with group-level features and thus lead to misleading results. Although it's true that the problem might be partially mitigated by tailoring traditional single-level analytical techniques such as multiple regression, it's also true that this somewhat ad hoc maneuver can simply bring about various new vexing and recalcitrant issues (Luke 2004; Nezlek 2008; Heck and Thomas 2015).

Second, the problem can also be viewed from the perspective of identifying explanatory variables. In building a mixed-effects model, the main consideration is often to find a couple of variables that may play the role of explaining the pattern or phenomenon observed in the data. Here a modeler must be clear about how to assign explanatory variables, for instance, she must consider if there are different levels of analyses and, if so, which explanatory variables should be assigned to what levels, and so on. These considerations may come before her model building because of background knowledge, which paves the way for her to develop a conceptual framework for investigating the problem of interest. However, without such a clear and rigorous consideration of identifying and assigning multilevel explanatory variables, an analysis can flaw simply because it confounds variables at different levels.

Respecting the multilevel nature of explanatory variables has another advantage: "Through examining the variation in outcomes that exists at different levels of the data hierarchy, we can develop more refined theories about how explanatory variables at each level contribute to variation in outcomes" (Heck and Thomas 2015, 33). In other words, in respecting the multilevel nature of

explanatory variables, we get a clear idea of how, and to what degrees, explanatory variables at different levels contribute to variation in outcomes. If these variables do contribute to variation in outcomes, as it always happens in MEM, then the situation suggests an image of *explanatory indispensability*: all the explanatory variables at different levels are indispensable to explaining the pattern or phenomenon of interest.

Given these considerations, therefore, one implication for the explanatory reductionism debate becomes clear: it isn't always the case that, given a relatively higher-level phenomenon it can be reductively explained by a relatively lower-level feature. Rather, in cases where the data show a nested structure or, put differently, the phenomenon suggests multilevel explanatory variables, we routinely combine the higher-level with the lower-level in a single (explanatory) model. As a result, one fundamental tenet of explanatory reductionism is violated: single level preference.

4.2. Interactions between levels

Another crucial feature of multilevel modeling is its emphasis on *a cross-level interaction*, which is defined as

> "The potential effects variables at one level of a data hierarchy have on relationships at another level […]. Hence, the presence of a cross-level interaction implies that the magnitude of a relationship observed within

> groups is dependent on contextual or organizational features defined by higher-level units". (Heck and Thomas 2015, 42-43)

Remember that there is a term $\gamma_{11} X_{ij} W_j$ in our mixed-effects model discussed in Section 3, which indicates the cross-level interaction between the group-level and the individual-level. More specifically, this term can be best construed as the impact of a group-level variable, e.g., group size, upon the individual-level relationship between a predictor, e.g., study time, and the outcome, e.g., students' scores.

The cross-level interaction points to the plain fact that an organization or a system can somehow influence its members or components by constraining how they behave within the organization or system. This doesn't necessarily imply top-down causation (Section 5.3 will turn back to this point). Within the context of scientific explanation, however, it does imply that it isn't simply that characteristics at different levels separately contribute to variation in outcomes, but rather that they interact in producing variation in outcomes. In other words, the pattern or phenomenon to be explained can be understood as generated by the interaction between explanatory variables at different levels. Therefore, to properly explain the phenomenon of interest, we need not only have a clear idea of how to assign explanatory variables to different levels but also an unequivocal conception of whether these explanatory variables may interact.

Different models can be built depending on different considerations of the cross-level interaction. To see this, consider the student-achievement-at-school

example again. In some experiment setting we may assume that there was no cross-level interaction between group-level characteristics and the individual-level relationship (between study time and scores). In such a situation, we kept the effect of individual study time on scores the same across different classes, i.e., we kept the slope constant across classes. In the meanwhile, we treated another group-level variable (i.e., intercept) as varying across classes, i.e., different classes have different average scores. So, this is a case where we have a clear idea of how to assign explanatory variables but no consideration of the cross-level interaction. Nonetheless, in a different experiment setting we may assume that there existed cross-level interaction, and hence the effect of individual study time on scores can no longer be kept constant across different classes. At the same time, we treated another group-level variable (i.e., intercept) as varying across classes. Hence, this is a case where we have both a clear idea of how to assign explanatory variables and a consideration of the cross-level interaction. Corresponding to these two different scenarios, two different mixed-effects models can be built, as shown below:

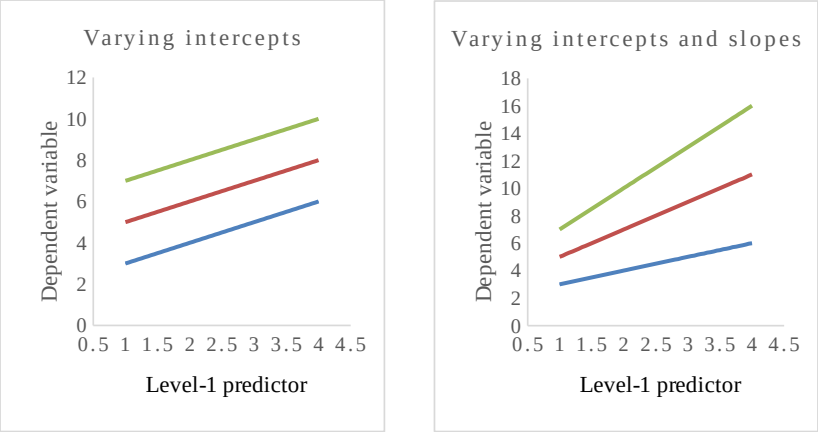


Figure 1. Two different models showing varying intercepts or varying intercepts and slopes, respectively. Three lines represent three classes. This figure is adapted from Luke (2004, 12).

Given such a cross-level interaction, therefore, the explanatory reductionist position has been further challenged. This is because any reductive explanation that privileges one level of analysis—usually the lower-level—over the others falls short of capturing this kind of interaction between levels. If they fail to do so, then they are missing important terms relevant to explaining the phenomenon of interest. As a consequence, a mixed-effects model involving interactions between levels simultaneously violates the two fundamental pillars of explanatory reductionism: first, it violates single level preference because it involves multilevel explanatory variables in explaining phenomena, and second, it violates lower-level obsession because it privileges no levels—all levels are interactively engaged in producing outcomes.

5. Potential Objections

This section considers two potential objections.

5.1. In-principle argument

One argument that resurfaces all the time in the reductionism-versus-antireductionism debate is the in-principle argument, the core of which is that even if reductive explanations in a field of study are not available for the time being, it doesn't follow that we won't obtain them someday (e.g., Sober 1999; Rosenberg 2006). Therefore, according to some reductionists, the gap between current-science and future-science is simply a matter of time, for advancement in techniques, experimentation and data collecting can surely fill in the gap.

However, I think the argument flaws. To begin with, advancement in techniques, experimentation and data collecting isn't always followed by reductive explanations. For example, in our MEM discussed in Section 3, even if the data about the individual-level is available and sufficiently detailed, it isn't the case that we explain the phenomenon of interest in terms of the data from the individual-level alone. Consider another example: in dealing with problems associated with complex systems in systems biology, even though large-scale experimentation (e.g., via computational simulation) can be conducted and high throughput data arranging over multiple scales/levels can be collected, a bottom-up reductive approach must be integrated with a top-down perspective so as to

produce useful explanations or predictions (Green 2013; Green and Batterman 2017; Gross and Green 2017).

Nevertheless, reductionists may reply that the situations presented above only constitute an in-practice impediment, for it doesn't undermine the *possibility* that lower-level reductive explanations, typically provided by some form of 'final science', will be available someday. Let us dwell on the notion of possibility a bit longer. The possibility here may be construed as a *logical possibility* (Green and Batterman 2017, 21; see also Batterman 2017). Nonetheless, if it's merely logically possible that there will be some final science providing only reductive explanations, then nothing can exclude another logical possibility that there will be some 'mixed-science' providing only multilevel explanations. After all, how can we decide which logical possibility is more possible (or logically more possible)? I doubt that logic alone could provide anything useful in justifying which possibility is more possible, and that appealing to logical possibility could offer anything insightful in helping us understand how science proceeds. As Batterman puts, "Appeals to the possibility of *in principle* derivations rarely, if ever, come with even the slightest suggestion about how the derivations are supposed to go" (2017, 12; author's emphasis).

Another interpretation of possibility may be associated with real possibilities, referring to the actual cases of reductive explanations happening in science. Unfortunately, I don't think the real scenario in science speaks for the reductionist under this interpretation. Though it's impossible to calculate the absolute cases of non-reductive explanations occurring in science, a cursive look at scientific

practice can tell that a large portion of scientific explanations proceeds in a non-reductive fashion, as suggested by multilevel modeling (Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017; Mitchell and Gronenborn 2017). Moreover, even in areas such as physics which was regarded as a paradigm for the reductionist stance, progressive explanatory reduction doesn't always happen (Green and Batterman 2017; Batterman 2017).

In sum, we have shown that the in-principle argument fails for it neither offers help in understanding how science proceeds if it's construed as implying a logical possibility, nor goes in tune with scientific practice if it's construed as implying real possibilities.

5.2. Top-down causation

In Section 3 we have shown that there is a cross-level interaction taking the form that higher-level features may impact lower-level features. A worry arises: Does this imply top-down causation?

My answer to this question is twofold. First, it's clear that this short essay isn't aimed to engage in the philosophical debate about whether, and in what sense, there exists top-down causation (see Craver and Bechtel 2007; Kaiser 2015; Bechtel 2017). Second, what we can do now is to show that the cross-level interaction is a clear and well-defined concept in multilevel modeling. It unambiguously means the constraints on the lower-level processes exerted by the higher-level parameters (Green and Batterman 2017). In our multilevel modeling

discussed in Section 3, we have shown that group-level features may impact some individual-level features through the way that each group possesses its own feature relevant to explaining the differences at the individual-level across groups. This idea is incorporated into the mixed-effects model by assigning some explanatory variables to the group-level and a cross-level interaction term to the model.

The idea of cross-level-interaction-as-constraint is widely accepted in multilevel modeling broadly construed, where constraint is usually expressed in the form of initial and/or boundary conditions. For example, in modeling cardiac rhythms, due to "the influences of initial and boundary conditions on the solutions of the differential equations used to represent the lower level process" (Noble 2012, 55; Cf. Green and Batterman 2017, 32), a model cannot simply narrowly focus on the level of proteins and DNA but must also consider the levels of cell and tissue working as constraints. The same story happens in cancer research, where scientists are advocating the idea that tumor development can be better understood if we consider the varying constraints exerted by tissue (Nelson and Bissel 2006; Shawky and Davidson 2015; Cf. Green and Batterman 2017, 32).

6. conclusion

This essay has shown that no-reductive explanations involving many levels predominate in areas where the systems under consideration exhibit a hierarchical structure. These explanations violate the fundamental pillars of explanatory

reductionism: single level preference and lower-level obsession. Traditional single-level reductive approaches fall short of capturing systems of this kind because they face the challenges of committing either the atomistic or ecological fallacy.

References

Batterman, Robert. 2013. The "Tyranny of Scales." In *The Oxford Handbook of Philosophy of Physics*, ed. Robert Batterman, 255-286. Oxford: Oxford University Press.

———. 2017. "Autonomy of Theories: An Explanatory Problem." *Noûs* 1-16.

Bechtel, William. 2010. "The Downs and Ups of Mechanistic Research: Circadian Rhythm Research as an Exemplar." *Erkenntnis* 73:313–328.

———. 2017. "Explicating Top-Down Causation Using Networks and Dynamics." *Philosophy of Science* 84:253–274.

Bickle, John. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Brigandt, Ingo. 2010. "Beyond Reductionism and Pluralism: Toward an Epistemology of Explanatory Integration in Biology." *Erkenntnis* 73 (3): 295-311.

———. 2013a. "Explanation in Biology: Reduction, Pluralism, and Explanatory Aims." *Science and Education* 22:69–91.

———. 2013b. "Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:461–465.

Craver, Carl. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, Carl, and William Bechtel. 2007. “Top-down Causation without Top-Down Causes.” *Biology and Philosophy* 22:547–563.

Freedman, David. 1999. “Ecological Inference and the Ecological Fallacy.” In *International Encyclopedia of the Social and Behavioral* Sciences, vol. 6, ed. Neil Smelser, and Paul Baltes, 4027–4030. New York: Elsevier.

Green, Sara. 2013. “When One Model Isn’t Enough: Combining Epistemic Tools in Systems Biology.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:170–180.

Green, Sara, and Robert Batterman. 2017. “Biology Meets Physics: Reductionism and Multi-Scale Modeling of Morphogenesis.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 61:20–34.

Gross, Fridolin, and Sara Green. 2017. “The Sum of the Parts: Large-Scale Modeling in Systems Biology.” *Philosophy, Theory, and Practice in Biology* 9: (10).

Heck, Ronald, and Scott Thomas. 2015. *An Introduction to Multilevel Modeling Techniques* (3rd Edition). New York: Routledge.

Hull, David. 1972. “Reductionism in Genetics—Biology or Philosophy?” *Philosophy of Science* 39 (4): 491-499.

Hüttemann, Andreas, and Alan Love. 2011. “Aspects of Reductive Explanation in Biological Science: Intrinsicality, Fundamentality, and Temporality.” *British Journal for the Philosophy of Science* 62 (3): 519-549.

Kaiser, Marie. 2015. *Reductive Explanation in the Biological Sciences*. Springer.

Knuuttila, Tarja. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Studies in History and Philosophy of Science Part A* 42:262–271.

Luke, Douglas. 2004. *Multilevel Modeling*. London: SAGE Publications, Inc.

Maxwell, Sophie, Katherine Reynolds, Eunro Lee, et al. 2017. “The Impact of School Climate and School Identification on Academic Achievement: Multilevel Modeling with Student and Teacher Data.” *Frontiers in Psychology* 8:2069.

Mitchell, Sandra. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

———. 2009. *Unsimple Truths: Science, Complexity, and Policy*. Chicago: University of Chicago Press.

Nezlek, John. 2008. “An Introduction to Multilevel Modeling for Social and Personality Psychology.” *Social and Personality Psychology Compass* 2/2 (2008):842–860.

Noble, Daniel. 2012. “A Theory of Biological Relativity: No Privileged Level of Causation.” *Interface Focus* 2(1):55–64.

O’Malley Malley, Ingo Brigandt, Alan Love, et al. 2014. “Multilevel Research Strategies and Biological Systems.” *Philosophy of Science* 81:811–828.

Rosenberg, Alex. 2006. *Darwinian Reductionism, or How to Stop Worrying and Love Molecular Biology*. Chicago: University of Chicago Press.

Sarkar, Sahotra. 1998. *Genetics and Reductionism*. Cambridge: Cambridge University Press.

Schagen, I. P. 1990. “Analysis of the Effects of School Variables Using Multilevel Models.” *Educational Studies* 16:61–73.

Shawky, Joseph, and Lance Davidson. 2015. “Tissue Mechanics and Adhesion during Embryo Development.” *Developmental Biology* 401(1):152–164.

Sober, Elliot. 1999. “The Multiple Realizability Argument against Reductionism.” *Philosophy of science* 66:542–564.

Wang, Yau-De, and Hui-Hsien Hsieh. 2012. “Toward a Better Understanding of the Link Between Ethical Climate and Job Satisfaction: A Multilevel Analysis.” *Journal of Business Ethics* 105:535–545.

Waters, C. Kenneth. 2008. “Beyond Theoretical Reduction and Layer-Cake Antireduction: How DNA Retooled Genetics and Transformed Biological Practice”. In *The Oxford Handbook of Philosophy of Biology*, ed. Michael Ruse, 238-262. New York: Oxford University Press.

Weber, Marcel. 2005. *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

The Universe Never Had a Chance

C. D. McCoy[‡]

1 March 2018

Abstract

Demarest asserts that we have good evidence for the existence and nature of an initial chance event for the universe. I claim that we have no such evidence and no knowledge of its supposed nature. Against relevant comparison classes her initial chance account is no better, and in some ways worse, than its alternatives.
Word Count: 4712

1 Introduction

Although cosmology, the study of the universe's evolution, has largely become a province of physics, philosophical speculation concerning cosmogony, the study of the origin of the universe, continues up to the present. Certainly, many believe that science has settled this too by way of the well-known and well-confirmed big bang model of the universe. According to the big bang account the universe began in a extremely hot, dense state, composed of all the different manifestations of energy that we know. Indeed, time itself began with the big bang. Yet, properly speaking, the universe's past singularity is not some event in spacetime according to the general theory of relativity. In cosmological models this hot dense state called the big bang is generally understood instead as just a very early stage of the universe's evolution, i.e. properly a part of cosmology and not cosmogony. While we may be highly confident that the entire big bang story is correct back to a very early time, our confidence should at some point decrease as we near the supposed "first moment". Thus there remains world enough and time to engage in traditional philosophical and scientific speculations about cosmogony and cosmology alike. Were there previous stages to the universe? What brought the universe into existence? What was the character of this initial happening (should it in fact exist)?

The ubiquity of probabilities in modern physical theories, e.g. quantum mechanics and statistical mechanics, has led some to wonder as well how chance should fit into our

[*]**Acknowledgements:** Pending.

[†]School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK. email: casey.mccoy@ed.ac.uk

cosmogonical worldview. In this vein, Demarest (2016) argues that the probabilities of all events in a(n ostensibly) deterministic universe can be derived from an initial chance event and, what's more, that "we have good evidence of its existence and nature." In this paper I aim to dispute these latter claims. I argue that we do not have any evidence at all of an initial chance event in a big bang universe as described above, much less of its nature. What we rather have in Demarest's account is just a particular way of interpreting probabilistic theories, where all probabilities are taken to derive from ontic chances pertaining to the particular genesis of the relevant physical system, e.g. the universe as a whole. I claim that this interpretation, while coherent, should be disfavored in cosmology—we should rather say that *the universe never had a chance*.[1] Along the way I will make several clarifying remarks concerning the relation of chance and determinism, cosmological probabilities, and alternative interpretations of statistical and quantum mechanics.

2 Chance and Determinism in Physical Theory

By the *world* metaphysicians usually mean something like "the maximally inclusive entity whose parts are all the things that exist." Of course terminology varies. This particular rendering comes from Schaffer (2010, 33), who instead chooses to call this entity the *cosmos*. Cosmologists do not usually call their object of study the cosmos; more commonly they say that they study the *universe*. In *Cosmology: The Science of the Universe*, Harrison explicitly notes the philosophical and historical dimensions of the world taken in its broadest sense, designating this world as a whole the *Universe*. Cosmology, according to Harrison, is the study of universes, by which he means particular models of the Universe (Harrison, 2000, Ch. 1). Cosmological models are the particular concern of physical cosmologists; they are physical models of the Universe, which describe especially its large-scale structure and the evolution thereof.

In what follows I employ these terminologies in the following way. By the *world* I designate the locus of (principally) metaphysical questions concerning the Universe. Is the world deterministic? Is it chancy? By the *universe* I designate the locus of principally physical questions concerning the Universe. How did the big bang universe begin? How will it end? These are questions to which the big bang model should provide an answer.

I do not mean, of course, to introduce an admittedly arbitrary distinction between science and metaphysics by differentiating universes and worlds. Indeed, when one asks whether the world is deterministic, many metaphysicians of science would look first to models of the Universe to help decide the question. Wüthrich for example remarks, matter-of-factly, that "this metaphysical question deflates into the question of whether our best physical *theories* entail that the world is deterministic or indeterministic" (Wüthrich, 2011, 366).

[1] There are several senses, in fact, in which this claim is true. Cosmology suggests that the inevitable fate of the universe is to become ever more sparse and empty through the accelerated expansion of space under the influence of dark energy.

Indeed, many discussions of determinism adopt the approach mentioned by Wüthrich. Let *determinism* denote the thesis that the world is deterministic. Then, following for example (Lewis, 1983, 360), a world is *deterministic* if and only if the laws of that world are deterministic. To determine whether the laws of the universe are deterministic, we must look to our theories of which those laws are part and ask whether those laws taken together should be considered deterministic. It is by no means a straightforward matter to decide whether a given physical theory is deterministic of course. Even the classic example of deterministic physics, Newtonian mechanics, admits many counterexamples against its putative determinism (Earman, 1986; Norton, 2008). General relativity as well seemingly permits indeterministic phenomena in the form of causal pathologies (closed timelike curves) (Earman, 1995) and, if the hole argument is to be believed, is hopelessly rife with indeterminism (Earman and Norton, 1987).

Although classical theories like classical mechanics and general relativity are nevertheless debatably deterministic, surely probabilistic theories like quantum mechanics are properly characterized as indeterministic (at least so long as the probabilities involved are objective features of the world). Yet various interpretations of probabilistic theories seek to avoid indeterminism even here, where it seems unassailable, by characterizing probabilities as merely epistemic or subjective, or else by presenting them as fully deterministic theories (as in the Bohmian interpretation of quantum mechanics). Philosophers have raised serious concerns, however, over how one can truly understand probabilities in deterministic theories, an issue that has been termed the "paradox of deterministic probabilities" (Loewer, 2001; Winsberg, 2008; Lyon, 2011) in statistical mechanics, since objective probabilities seem to entail indeterminism necessarily.

The most well-known and successful reconciliation of chance and determinism in the context of statistical mechanics is defended by Loewer (2001). It is seldom recognized by interpreters, however, that there is no reconciliation in the sense of simultaneous compatibility between chance and determinism. The world cannot both be chancy and deterministic as a matter of metaphysical fact. As Lewis writes, "to the question of how chance can be reconciled with determinism, or to the question of how disparate chances can be reconciled with one another, my answer is: *it can't be done* (Lewis, 1986, 118). This is because chance entails indeterminism, the contrary of determinism. Thus, insofar as the probabilities of statistical mechanics and quantum mechanics are objective, these theories are indeterministic theories. Loewer's account actually shows us how deterministic laws can co-exist with indeterministic laws within a theory. The source of all probabilities in statistical mechanics, according to Loewer, is in an initial chance distribution over microscopic states of affairs. After the initial time these states of affairs evolve deterministically. Note that although for almost all times evolution is deterministic, it is not so at all times. There is an initial chance event, which is where the indeterminism of the theory appears. A deterministic theory is, recall, a theory whose laws are deterministic, not a theory whose laws are mostly deterministic or operate deterministically for almost all times.

Loewer's account is also presented in terms of Humean chances, so he does not believe

these chances and laws actually exist. According to the modern Humean, they merely are the result of the best systematizations of the occurrent facts, in keeping with Lewis's "best systems account" of laws and chances. Demarest, however, offers a small tweak to Loewer's Humean account by invoking a "robustly metaphysical account of chance" (Demarest, 2016, 256). She claims that such chances are compatible with determinism, and indeed they are when, as said, compatibility is understood to pertain to the co-existence of indeterministic and deterministic laws in a single theory—which, however, do not operate at the same time.[2]

Demarest's central claims are that this initial chance event exists and that we have good evidence for it. I dispute these claims in the remainder of the paper.

To begin, it is not so clear what exactly Demarest takes the evidence for the initial chance event to be. She does contrast the evidential position of her view with the Humean view of Loewer, claiming that, "for the Humean, the statistical patterns in the world are not evidence of an initial chance event" (Demarest, 2016, 261)—presumably this is so because Humeans reject the metaphysics of chance for the usual Humean reasons. One might suppose, then, that she believes that statistical patterns in the world are evidence of an initial chance event for all those who do not share the Humeans ontological worries. Let us accept, for the moment then, that statistical patterns may be *some* evidence for the existence of chances, for it is difficult to see what other evidence there might be for an initial chance event. In that case, on what grounds might we say that statistical patterns are good evidence for initial chances? I consider a series of three salient contrast classes.

First, do statistical patterns in data provide good evidence for indeterministic (i.e. chancy) theories *rather than deterministic theories*? It would seem that the answer is: not necessarily. (Werndl, 2009), for example, argues for the observational equivalence of indeterministic theories and deterministic theories. If one could contrive a fully deterministic theory that reproduces the same statistical patterns of the relevant phenomena observed in nature, then it would seem that such patterns provide no better evidence for the indeterministic theory than the deterministic one. However, since the theories under discussion, statistical mechanics and quantum mechanics, are generally characterized as indeterministic, let us flag but set aside the possibility of fully deterministic alternatives to them.

So, second, do statistical patterns provide good evidence for initial chances *rather than non-initial chances*? It would seem that the answer is firmly: no. There is a variety of ways one could implement chances into a probabilistic theory like statistical mechanics. All one must do, as Loewer shows us by example, is neatly separate when the indeterministic laws are operative and when the deterministic laws are operative. Loewer chooses to locate all the indeterminism in one place—the initial time—but one could equally locate it at another time, at many times, or even all times. Statistical mechanics does not wear its interpretation on its sleeve, just as quantum mechanics does not decide between solutions of the measurement problem, whether initial chances as in Bohmian mechanics or collapse

[2]Still, it is worth emphasizing that her claim that her account applies to deterministic worlds is false, for chancy worlds are not deterministic.

dynamics as in GRW (discrete time collapses) or CSL (continuous collapses). Unless there are evidential reasons to favor one implementation of indeteterministic probabilities over the others, there is not good evidence for an initial chance event. Certainly statistical patterns in nature will not do so.

Third, do statistical patterns provide good evidence for "robustly metaphysics" chances *rather than Humean chances*? It seems as if this might Demarest's intended contrast class, since much of the discussion in the paper concerns the Humean account. I will have something to say about the relative merits of Demarest's non-Humean account and Loewer's Humean account at the end of the next section. In any case though, it does not seem as if statistical patterns decide the matter in Demarest's mind, for she repeatedly demurs in the face of Humean responses to the considerations she raises, claiming only to offer an alternative "for philosophers who are antecedently sympathetic to governing laws of nature or powerful properties" (Demarest, 2016, 261-2). She finds it "plausible to think of the universe as having an initial state and as producing subsequent states in accordance with the laws of nature (some of which may be chancy)" (Demarest, 2016, 261). Such metaphysical intuitions are not grounded on observations of statistical patterns. Statistical patterns do not have any evidential bearing on the metaphysical dispute between the Humean and non-Humean.

Therefore, based on my canvassing of relevant alternatives, I conclude that we in fact do not have good evidence for an initial chance event, where evidence is interpreted in terms of statistical patterns (or in any usual sense of the term "evidence"). At best we have a motivation to attend to indeterministic theories when our evidence displays statistical patterns. It is another matter entirely to decide how to implement probabilities in that theory.

That said, Demarest's reasoning could be interpreted at points as invoking explanatory considerations as justification for the initial chance interpretation. Insofar as one considers "what justifies" as constituting evidence, perhaps these explanatory considerations should be counted as evidence.[3] Nevertheless, it does not look, on the face of it, like we have good evidence for an initial chance event still. Repeating the three cases considered before: deterministic and chancy theories can both serviceably explain statistical evidence; alternative implementations of chance in interpretations of indeterministic theories explain statistical evidence equally well; Humean and non-Humean metaphysics each render a story for how statistical patterns come about (merely subjective intuitions notwithstanding). Without explicit explanatory reasons to prefer one of these alternatives to the other, reasons lacking in Demarest's argument, good evidence (in this wider sense) for an initial chance event remains elusive.

[3]There are obvious dangers with going to far in this direction. Suppose that the Supreme Being explains all. Then it would appear that we have very good evidence of Its existence, which is obviously absurd.

3 Chance and Determinism in Systems of the World

In the previous section I gave reasons to doubt Demarest's claims about an initial chance event and our evidence for it. I disputed especially that we have evidence for it and did so by comparing it to alternatives of three different kinds. In the first case I characterized the issue (in part) as a matter of theory choice, namely of choosing between an indeterministic and deterministic theory. In the second case I characterized the issue as a matter of theory interpretation, namely of interpreting between different ways of implementing probability in a theory that does not decide one way or another on how this must be done. In the third case I characterized the issue as a matter of metaphysics, namely of deciding between the ontological status of chances.

In this section I consider more broadly whether there are any reasons to favor Demarest's interpretation, in particular in the sense of the just given second characterization of the issue. The question is whether the world should be thought to have an initial chance event, when one might consider that it is chancy in various other ways, e.g. its laws of evolution themselves are always probabilistically indeterministic.

First of all, it is worth mentioning that from the point of view given by the contemporary standard model of cosmology this question is moot. The so-called ΛCDM model, a development of the older standard big bang model, is a model of the general theory of relativity, a theory which makes use of no probabilities at all in its basic description of gravitating systems (including the universe). In this different sense it is also true that the universe never had a chance.

Demarest is not particularly interested in cosmology or the universes of general relativity however. She is concerned with probabilistic theories like classical statistical mechanics and quantum mechanics as applied to the world at large. We should, that is, imagine a statistical mechanical universe or a quantum mechanical universe (never minding that no concrete such model exists in physics that describes our universe) as a conceptual possibility when asking metaphysical questions about the world. Given the different ways of implementing probabilities in such a universe, we should ask whether one way is preferable to the others.

I should point out that this is not Demarest's question, for she explicitly restricts attention to "deterministically evolving worlds". Of course these worlds are not actually deterministic so long as the probabilities involved are chances. Nevertheless, unaffected by that fact is one of her central points: "that positing just one initial chance event can justify the usefulness and explain the ubiquity of nontrivial probabilities to epistemic agents like us, even if there are no longer any chance events in our world"(Demarest, 2016, 249). I say: so can a lot of other ways of conceiving chance in these theories. It is therefore necessary to compare them if we are to take Demarest's (and Loewer's) account seriously.

For present purposes, I am happy to agree with Demarest that the initial chance account can indeed justify and explain nontrivial probabilities used to describe subsystems of the universe.[4] But is it a good explanation? Is it worth believing?

[4]Notwithstanding pressure to move in this "global" direction in statistical mechanics (Callender, 2011)

The initial chance account invites the oft-invoked (in cosmology) picture of the (blind and unskilled) Creator throwing a dart (Wald, 2006, 396) or pointing a pin (Penrose, 1989, 442) at the set of possible universes, thereby picking out the initial conditions of the universe. That such pictures are intended as pejorative jabs at dubious metaphysics is plain. A mere picture is hardly an objection, of course, so what is it that seems problematic about initial chances for the universe? Could it not be the best cosmogonical story of our universe, that is, that a matter of chance determined its actualization out of a vast range of possibilities that could have been actualized had only their sisal been struck?

Intuition suggests that this just is not a serious, satisfying story for how the world could be. The probabilities of events in the actual world would derive ultimately from the probabilities for the actualization of our world. But why should we not just assume that the world started in the state that it did, with probability one or with certainty? Presumably the response of the initial chance advocate is that in that case we would lose the justification and explanation of subsystem probabilities. Yet is there anything to lose, if this metaphysical explanation is epistemically untrustworthy? How can we come to know these ultimate probabilities of other worlds? Is the metaphysical story sufficiently complete even? How could the probabilities of other worlds matter for what happens in *our* world?

I am willing to grant that these questions do have some answer, for what strikes me as a more serious difficulty is the following. Insofar as they are objective and justified, the probabilities agents like us use for specific events in subsystems of the world must be epistemic probabilities. On Demarest's (and Loewer's) account all such epistemic probabilities derive from initial epistemic probabilities for different initial conditions of the world. How is it that these probabilities obtain their needed objectivity and justification, and hence explanatory power? According to Demarest it is because they accord with the actual chances. However, what has one achieved by invoking "actual chances" at this stage? Although these chances do not merely have a *virtus dormitiva* per se, "just so" stories like this surely make the explanatory credentials of chances suspect. Does one dare invoke a transcendental argument or thump the realist table to defend their objectivity?

If we were somehow forced to adopt the initial chance explanation of epistemic probabilities, then we might swallow whatever dubious metaphysics attendant to it. If there were reasonable alternatives, however, should we not prefer them? And indeed there are other interpretive options available. Locating the chances at another time (or even "outside the universe") constitutes one set of possibilities, but they obviously suffer from the same awkwardness as the initial chance account. Another is based on the idea that chancy behavior occurs at discrete time intervals. One finds this idea in the orthodox Copenhagen and other collapse interpretations of quantum mechanics for example. One might be uneasy with the invocation of chancy behavior at potentially ill-defined times in such interpretations, and even with their postulation of two dynamical laws of nature, a deterministic one and an indeterministic one (although it is a feature of the initial chance account as well). However one at least avoids a commitment to chance figuring into

(and quantum mechanics) in order to justify and explain probabilities in subsystems of the universe, serious reservations about whether doing so is itself justified are advanced by, inter alia, Earman (2006).

cosmogenesis and also the questionable leap to objectivity in agential probabilities, since chances in these interpretations are physical processes that happen within the universe, whether as part of the general evolution of the universe or tied to the evolution of individual systems.

Another possibility is suggested by continuing this line of thought, i.e. of spreading chanciness out further in time. Instead of chancy behavior at discrete intervals, why not suppose that it occurs continuously? In quantum mechanics this idea is implemented in some interpretations, such as continuous spontaneous localization, and in statistical mechanics there are various stochastic dynamics approaches. Advantages of this idea are that one has a single law of evolution, an indeterministic one, and, again, one does not make chanciness a matter of cosmogenesis. What disadvantage? To some that it makes the world rife with indeterminism. Yet who is afraid of indeterminism? It surely does not mean anything goes, nor does it threaten the possibility of knowledge of the world (although there are limits to what we can know). Besides, by accepting quantum mechanics (or even statistical mechanics) we have already let indeterminism in the door in physics.

When we look at the interpretations available for a world governed by probabilistic laws, in every case the alternatives to the initial chances view therefore appear preferable. Indeed, it would seem that only one who demands that the world be as deterministic as possible could favor the initial chances view, but it is hard to see what motivation there could be for that demand. I therefore conclude, in a final sense, that *the universe never had a chance*.

That said, I emphasize that this judgment applies only to the case where we treat the universe as a statistical mechanical system or quantum mechanical system. In other words, the world is the universe, our world-metaphysics is our universe-metaphysics. The considerations leading to this conclusion change shape somewhat when we confine the application of our theories to systems describable by those theories. The initial chance account is far less dubious when attached to individual statistical mechanical systems and not automatically to the universe at large. Indeed, it could well be that the initial conditions of similar systems are best treated as randomly distributed, for here we do have empirical evidence that this interpretation can be used to explain—unlike with the universe, where we have but one system.

There is, as noted, sometimes pressure to globalize our theories, especially in the case of statistical mechanics. If we ask what accounts for the randomness in initial conditions of a particular class of systems, it is natural to look at larger systems that contain them. If we find that these systems have random initial conditions, then we continue to expand our scope, ultimately reaching the "maximally inclusive entity whose parts are all the things that exist." This globalization of statistical mechanics is the kernel of the so-called imperialism of (Albert, 2000) and Loewer. If we are right to feel this pressure to interpret the world at large in the same terms as individual physical systems, then there is concomitant pressure to hold the same interpretive of chance in both cases. I have argued, however, that the intuitive considerations vary somewhat, at least with respect to the initial chance account. Is this reason to disfavor it in the case of individual systems? Or is our confidence in its applicability for individual systems sufficient to overcome any hesitation at

accepting it for the universe? My inclination is to answer "yes" and "no", but I offer no grounds for the preference here. I do believe that metaphysicians of science should care about considerations like this, however, having to do with the relation of subsystem and universe, for often enough what seems right in one context is questionable in the other.

I close this section with a brief comment on the relation of Loewer's and Demarest's accounts. As I argued above, empirical evidence and explanatory considerations do not favor one over the other, since they account for empirical evidence in essentially the same way. The central difference is whether chances are understood as reducible to other facts, hence not part of the fundamental ontology of the world, or as "robustly metaphysical", in which case they are. The problems Demarest mentions for the Humean view—past events may have nontrivial chances, the chance of an event depends on what one knows, worlds with identical frequencies cannot have different chances, etc.—are surely not problems when viewed properly through the Humean lens. However, whereas the problem I raise for the initial chance view, concerning the explanatory credentials and justification for the posit of initial chances, threatens Demarest's account, it will not worry the Humean of Loewer's stripe, for these initial chances do not exist for the Humean. Humean chances do not produce or generate any actual states of affairs. Of course one may raise the usual complaint against the Humean, that there is a circularity in the Humean account involving descriptions explaining themselves, and others besides. I do not care to enter into this debate here of course. I only wish to point out that my argument about how chance can fit into a cosmogonical worldview appears to give some reason to favor the Humean account in this particular context.

4 Conclusion

In this paper I considered whether we should think that the world had one chance, as claimed by Demarest. First I considered her claim that we have good evidence that an initial chance event occurred by contrasting it with relevant classes of alternatives. I argued that evidence neither favors a chancy theory over a chanceless theory, nor initial chances over other implementations of chances, nor metaphysically robust chances over Humean chances. I concluded, therefore, that we do not have good evidence to adopt the initial chance account.

I then considered whether there were other reasons to favor or disfavor the initial chance account. I argued that the dubious nature of worldly chances provides a strong impulse to look for other accounts that do not make chance a matter of cosmogenesis. The other implementations did not suffer from this defect, so I suggested that from a cosmogonical perspective they should be preferred. But the relation of the universe and its subsystems makes a demand to have a consistent interpretation. As the initial chance account looks favorable on the subsystem level (to many) and not on the universe's level (as I argued), there remains a significant metaphysical tension to be resolved.

References

Albert, D. (2000). *Time and Chance*. Cambridge, MA: Cambridge, MA: Harvard University Press.

Callender, C. (2011). The past histories of molecules. In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 83–113. Oxford: Oxford University Press.

Demarest, H. (2016). The universe had one chance. *Philosophy of Science 83*(2), 248–264.

Earman, J. (1986). *A Primer on Determinism*. Dordrecht: D. Reidel Publishing Company.

Earman, J. (1995). *Bangs, Crunches, Whimpers, and Shrieks*. Oxford: Oxford University Press.

Earman, J. (2006). The "past hypothesis": Not even false. *Studies in History and Philosophy of Modern Physics 37*, 399–430.

Earman, J. and J. Norton (1987). What price spacetime substantivalism? the hole story. *British Journal for the Philosophy of Science 38*, 515–525.

Harrison, E. (2000). *Cosmology: the science of the universe* (2nd ed.). Cambridge: Cambridge University Press.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Lewis, D. (1986). *Philosophical Papers*, Volume 2. Oxford: Oxford University Press.

Loewer, B. (2001). Determinism and chance. *Studies in History and Philosophy of Modern Physics 32*, 609–620.

Lyon, A. (2011). Deterministic probability: neither chance nor credence. *Synthese 182*, 413–432.

Norton, J. (2008). The dome: An unexpectedly simple failure of determinism. *Philosophy of Science 75*, 786–798.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

Schaffer, J. (2010). Monism: The priority of the whole. *The Philosophical Review 119*, 31–76.

Wald, R. (2006). The arrow of time and the initial conditions of the universe. *Studies in History and Philosophy of Modern Physics 37*, 394–398.

Werndl, C. (2009). Are deterministic descriptions and indeterministic descriptions observationally equivalent? *Studies in History and Philosophy of Modern Physics 40*, 232–242.

Winsberg, E. (2008). Laws and chances in statistical mechanics. *Studies in History and Philosophy of Modern Physics 39*, 872–888.

Wüthrich, C. (2011). Can the world be shown to be indeterministic after all? In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 365–389. Oxford: Oxford University Press.

Draft paper for the symposium *Mechanism Meets Big Data: Different Strategies for Machine Learning in Cancer Research* to be held at *PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association* (Seattle, WA; 1-4 Nov 2018).

MECHANISTIC MODELS AND THE EXPLANATORY LIMITS OF MACHINE LEARNING

Emanuele Ratti[1], University of Notre Dame

Ezequiel López-Rubio, Universidad Nacional de Educación a Distancia, University of Málaga

Abstract

We argue that mechanistic models elaborated by machine learning cannot be explanatory by discussing the relation between mechanistic models, explanation and the notion of intelligibility of models. We show that the ability of biologists to understand the model that they work with (i.e. intelligibility) severely constrains their capacity of turning the model into an explanatory model. The more a mechanistic model is complex (i.e. it includes an increasing number of components), the less explanatory it will be. Since machine learning increases its performances when more components are added, then it generates models which are not intelligible, and hence not explanatory.

1. INTRODUCTION

Due to its data-intensive turn, molecular biology is increasingly making use of machine learning (ML) methodologies. ML is the study of generalizable extraction of patterns from data sets starting from a problem. A problem here is defined as a given set of input variables, a set of outputs which have to be calculated, and a sample (previously input-output pairs already observed). ML calculates a quantitative relation between inputs and outputs in terms of a predictive model by learning from an already structured set of input-output pairs. ML is expected to increase its performances when the complexity of data sets increase, where complexity refers to the number of input variables and the number of samples. Due to this capacity to handle complexity, practitioners think that ML is potentially able to deal with biological systems at the macromolecular level, which are notoriously complex. The development of ML has been proven useful not just for the

[1] mnl.ratti@gmail.com

complexity of biological systems *per se*, but also because biologists now are able to generate an astonishingly amount of data. However, we claim that the ability of ML to deal with complex systems and big data comes at a price; *the more ML can model complex data sets, the less biologists will be able to explain phenomena in a mechanistic sense*.

The structure of the paper is as follows. In Section 2, we discuss mechanistic models in biology, and we emphasize a surprising connection between explanation and model complexity. By adapting de Regt's notion of pragmatic understanding (2017) in the present context, we claim that if a how-possibly mechanistic model can become explanatory, then it must be intelligible to the modeler (Section 2.2, 2.3 and 2.4). Intelligibility is the ability to perform precise and successful material manipulations on the basis of the information provided by the model about its components. The results of these manipulations are fundamental to recompose the causal structure of a mechanism out of a list of causally relevant entities. Like a recipe, the model must provide instructions to 'build' the phenomenon, and causal organization is fundamental in this respect. If a model is opaque to these organizational aspects, then no mechanistic explanations can be elaborated. By drawing on studies in cognitive psychology, we show that the more the number of components in a model increases (the more the model is complex), the less the model is intelligible, and hence the less an explanation can be elaborated.

Next, we briefly introduce ML (Section 3). As an example of ML application to biology, we analyze an algorithm called PARADIGM (Vaske et al 2010), which is used in biomedicine to predict clinical outcomes from molecular data (Section 3.1). This algorithm predicts the activities of genetic pathways from multiple genome-scale measurements on a single patient by integrating information on pathways from different databases. By discussing the technical aspects of this algorithm, we will show how the algorithm generates models which are more accurate as the number of variables included in the model increases. By variables, here we mean biological entities included in the model and the interactions between them, since those entities are modeled by variables in PARADIGM.

In Section 4 we will put together the results of Section 2 and 3. While performing complex localizations more accurately, we argue that an algorithm like PARADIGM makes mechanistic models so complex (in terms of the number of model components) that no explanation can be constructed. In other words, ML applied to molecular biology undermines biologists' explanatory abilities.

2. COMPLEXITY AND EXPLANATIONS IN BIOLOGY

The use of machine learning has important consequences for the explanatory dimension of molecular biology. Algorithms like PARADIGM, while providing increasingly accurate localizations, challenge the explanatory abilities of molecular biologists, especially if we assume the account of explanation of the so-called mechanistic philosophy (Craver and Darden 2013; Craver 2007; Glennan 2017). In order to see how, we need to introduce the notion of mechanistic explanation, and its connection with the notion of intelligibility (de Regt 2017).

2.1 Mechanistic explanations

Molecular biology's aim is to explain how phenomena are produced and/or maintained by the organization instantiated by macromolecules. Such explanations take the form of mechanistic descriptions of these dynamics. As Glennan (2017) succinctly emphasizes, mechanistic models (often in the form of diagrams complemented by linguistic descriptions) are vehicles for mechanistic explanations. Such explanations show how a phenomenon is produced/maintained and constituted by a mechanism – mechanistic models explain by explaining *how*. As Glennan and others have noticed, a mechanistic description of a phenomenon looks like what in historical narrative is called *causal narrative*, in the sense that it "describes sequences of events (which will typically be entities acting and interacting), and shows how their arrangement in space and time brought about some outcome" (Glennan 2017, p 83). The main idea is that we take a set of entities and activities to be causally relevant to a phenomenon, and we explain the phenomenon by showing how a sequence of events involving the interactions of the selected entities produces and/or maintains the explanandum. In epistemic terms, it is a

matter of showing a chain of inferences that holds between the components of a model (e.g. biological entities). Consider for instance the phenomenon of restriction in certain bacteria and archaea (Figure 1). This phenomenon has been explained in terms of certain entities (e.g. restriction and modification enzymes) and activities (e.g. methylation). Anytime a bacteriophage invades one of these bacteria or archaea (from now on *host cells*), host cells stimulate the production of two types of enzymes, i.e. a restriction enzyme and a modification enzyme. The restriction enzyme is designed to recognize and cut specific DNA sequences. Such sequences, for reasons we will not expose here[2], are to be found in the invading phages and/or viruses. Hence, the restriction enzyme destroys the invading entities by cutting their DNA. However, the restriction enzyme is not able to distinguish between the invading DNA and the DNA of the host cell. Here the modification enzyme helps, by methylating the DNA of the host cell at specific sequences (the same that the restriction enzyme cuts), thereby preventing the restriction enzyme to destroy the DNA of the host cell. The explanation of the phenomenon of restriction is in terms of a narrative explaining how certain entities and processes contribute to the production of the phenomenon under investigation. The inferences take place by thinking about the characteristics of the entities involved, and how the whole functioning of the system can be recomposed from entities themselves.

[2] See for instance (Ratti 2018)

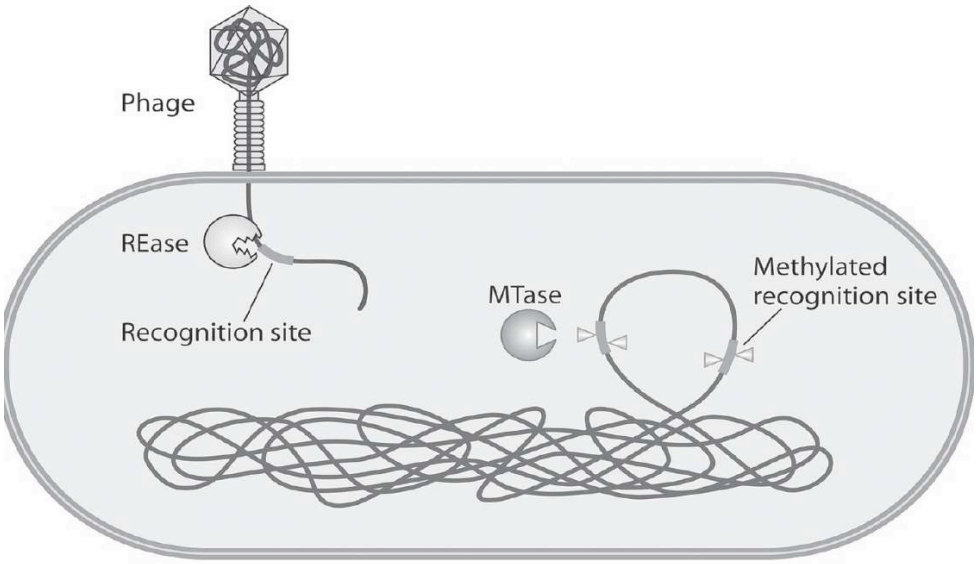


Figure 1. Mechanistic model of restriction. A phage enters a bacterium cell and sequences of its DNA are cleaved by a restriction enzyme (REase). Simultaneously, a modification enzyme (MTase) methylates a specific sequence in the DNA of host so that the restriction enzyme does not cleave the genome of the host too. Original figure taken from (Vasu and Nagaraja 2013).

2.2. Complexity of mechanistic models

Despite the voluminous literature on mechanistic explanation, there is a connection between models, *in fieri* explanations and the modeler that has not been properly characterized. In particular, mechanistic models should be intelligible to modelers in order to be turned into complete explanations. Craver noticed something like that when he states that his ideal of completeness of a mechanistic description (in terms of molecular details) should not be taken literary, but completeness always refer to the particular explanatory context one is considering. The reason why literary completeness is unattainable is because complete models will be of *no use* and completely *obscure* to modelers; "such descriptions would include so many potential factors that they would be *unwieldy for the purpose of prediction and control and utterly unilluminating to human beings*" (2006, p 360, emphasis added).

We rephrase Craver's intuitions by saying that *how-possibly models cannot be turned into adequate explanations if they are too complex*. We define complexity as *a function of the number of entities and activities (i.e. components of the model) that have*

to be coordinated in an organizational structure in the sense specified by mechanistic philosophers. This means that no agent can organize the entities and/or activities localized by highly complex models in a narration that rightly depicts the organizational structure of the *explanandum*. Therefore, very complex models which are very good in localization cannot be easily turned into explanations. Let us show why complex models cannot be turned into explanatory models in the mechanistic context.

2.3 Intelligibility of mechanistic models

The idea that agents cannot turn highly complex mechanistic models into explanations can be made more precise by appealing to the notion of *intelligibility* (de Regt 2017).

By following the framework of models as mediators (Morgan and Morrison 1999), de Regt argues that models are the way theories are applied to reality. Similar to Giere (2010), de Regt thinks that theories provide principles which are then articulated in the form of models to explain phenomena; "[t]he function of a model is to represent the target system in such a way that the theory can be applied to it" (2017, p 34). He assumes a broad meaning of explanation, in the sense that explanations are arguments, namely attempts to "answer the question of why a particular phenomenon occurs or a situation obtains (…) by presenting a systematic line of reasoning that connects it with other accepted items of knowledge" (2017, p 25). *Ça va sans dire*, arguments of the sort are not limited to linguistic items[3]. On this basis, de Regt's main thesis is that a *condition sine qua non* to elaborate an explanation is that the theory from which it is derived must be intelligible.

In de Regt's view, the intelligibility of a theory (*for scientists*) is "[t]he value that scientists attribute to the cluster of virtues (…) that facilitate the use of the theory for the construction of models" (p 593). This is because an important aspect of obtaining explanations is to derive models from theories, and to do that a scientist must use the theories. Therefore, if a theory possesses certain characteristics that make it easier to be used by a scientist, then the same scientist will be in principle more successful in deriving explanatory models. In (2015) de Regt extends this idea also to models in the sense that "understanding consists in being able to use and manipulate the model in order to make

[3] Mechanistic explanations are arguments, though not of a logical type

inferences about the system, to predict and control its behavior" (2015, p 3791). If for some reasons models and theories are not intelligible (to us), then we will not be able to develop an explanation, because we would not know how to use models or theories to elaborate one.

This idea of intelligibility of models and its tight connection with scientific explanation, can be straightforwardly extended to mechanistic models. Intelligibility of mechanistic models is defined by the way we *successfully* use them to explain phenomena. But how do we use models (mechanistic models in particular), and for what? Please keep in mind that whatever we do with mechanistic models, it is with explanatory aims in mind. Anything from predicting, manipulating, abstracting, etc is because we want an explanation. This is a view shared both by mechanistic philosophers but by de Regt as well, whose analysis of intelligibility is in explanatory terms.

First, highly abstract models can be used to build more specific models, as in the case of schema (Machamer et al 2000; Levy 2014). A schema is "a truncated abstract description of a mechanism that can be filled with descriptions of known component parts and activities" (Machamer et al 2000, p 16). For instance, consider the model of transcription. This model can be highly abstract where 'gene' stands for any gene, and 'transcription factor' stands for any transcription factor. However, we can instantiate such a schema in a particular experimental context by specifying which gene and which transcription factors are involved. The idea is that biologists, depending on the specific context they are operating, can instantiate experiments to find out which particular gene or transcription factor is involved in producing a phenomenon at a given time.

Next, mechanistic models can be used in the context of the *build-it test* (Craver and Darden 2013) with confirmatory goals in mind. Since mechanistic explanations may be understood as recipes for construction, and since recipes provide instructions to use a set of ingredients and instruments to produce something (e.g. a cake), then mechanistic models provide instructions to build a phenomenon or instructions to modify it in controlled ways because, after all, they tell us about the internal division of labor between entities causally relevant to producing or maintaining phenomena. This is in essence the build-it test as a confirmation tool; by modifying an experimental system on the basis of the 'instructions' provided by the model that allegedly explains such a phenomenon, we

get hints as to how the model is explanatory. If the hypothesized modifications produce in the 'real-world' the consequences we have predicted on the basis of the model, then the explanatory adequacy of the model is corroborated. The more the modifications suggested are precise, the more explanatory the model will be[4]. A first lesson we can draw is that *if a mechanistic model is explanatory, then it is also intelligible*, because it is included in the features of being explanatory mechanistically the fact that we can use the model to perform a build-it test.

The build-it test is also useful as a *tool to develop* explanations. Consider again the case of restriction in bacteria and a how-possibly model of this phenomenon based on a few observations. Let's say that we have noticed that when phages or viruses are unable to grow in specific bacteria, such bacteria also produce two types of enzymes. We know that the enzymes, the invading phages/viruses and restriction are correlated. The basic model will be as follows; anytime a phage or a virus invade a bacterium, these enzymes are produced, and hence the immune system of the bacterium must be related to these enzymes. We start then to instantiate experiments on the basis of this simple model. Such a model suggests that these enzymes must do something to the invading entities, but that somehow modify the host cell as well. Therefore, the build-it test would consist in a set of experiments to stimulate and/or inhibit these entities to develop our ideas about the nature of their causal relevance and their internal division of labor. *In fieri* mechanistic models suggest a range of instructions to 'build' or 'maintain' phenomena. These instructions are used to instantiate experiments to refine the model and make it explanatory. This is an example of what Bechtel and Richardson would call *complex localization* (2010, Chapter 6), and it is complex because the strategy used to explain the behavior of a system (immune system of host cells) is heavily constrained by empirical results of lower-levels. The how-possibly model affords a series of actions leading to a case of complex localization, when "constraints are imposed, whether empirical or theoretical, they can serve simultaneously to vindicate the initial localization and to develop it into a full-blooded mechanistic explanation" (Bechtel and Richardson 2010, p 125). Therefore, *if a how-possibly model can be turned into an explanatory model, then it*

[4] Please note that such a test, when involving adequate mechanistic explanations, is also the preferred way to teach students in text books, or also a way to provide instructions to reproduce the results of a peer-reviewed article

is intelligible, because the way we turn it into an explanatory model is by instantiating build-it tests.

A mechanistic model is therefore intelligible either when (a) it is a schema and we can instantiate such a model in specific contexts, or (b) when it affords a series of built-it test which are used either to corroborate its explanatory adequacy, or to make it explanatory. About (b), it should be noted that if we consider a mechanistic model as a narrative, then the model will be composed of a series of steps which influence each other in various ways. *Being able to use a model means being able to anticipate what would happen to other steps if I modify one step in particular*. This is not a yes/no thing. The model of restriction-modification systems is highly intelligible, because I know that if I prevent the production of modification enzymes I simultaneously realize that the restriction enzyme will destroy the DNA of the host cell. However, more detailed models will be less intelligible, because it would be difficult to simultaneously anticipate what would happen at each step by modifying a step in particular.

2.4 Recomposing mechanisms and intelligibility

In the mechanistic literature, the process of developing an explanatory model out of a catalogue of entities that are likely to be causally relevant to a phenomenon is called *recomposition of a mechanism* and it usually happens after a series of localization steps.

To recompose a mechanism, a modeler must be able to identify causally relevant entities and their internal division of labor. The idea is not just to 'divide up' a given phenomenon in tasks, but also a given task in subtasks interacting in the overall phenomenon, as it happens in complex localization (Bechtel and Richardson 2010). In the simplest case, researchers assume linear interactions between tasks, but there may be also non-linear or more complex type of interactions.

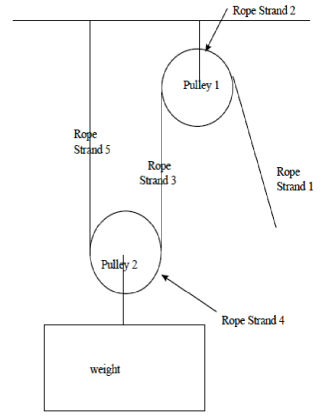
These reasoning strategies are usually implemented by thinking about these dynamics with the aid of *diagrams*. Diagrammatic representations usually involve boxes standing for entities (such as genes, proteins, etc) and arrows standing for processes of various sorts (phosphorylation, methylation, binding, releasing, etc). Therefore, biologists recompose mechanisms as mechanistic explanations by thinking about these diagrams,

and they instantiate experiments (i.e. built-it test) exactly on the basis of such diagrammatic reasoning.

Cognitive psychology and studies of scientific cognition have extensively investigated the processes of diagrammatic reasoning (Hegarty 2000; 2004; Nersessian 2008). Moreover, empirical studies have emphasized the role of diagrams in learning and reasoning in molecular biology (Kindfield 1998; Trujillo 2015). In these studies, diagrammatic reasoning is understood as a "task that involves inferring the behavior of a mechanical system from a visual-spatial representation" (Hegarty 2000, p 194). Hegarty refers to this process as *mental animation*, while Nersessian (2008) thinks about this as an instantiation of *mental modelling*. This is analogous to thinking about mechanistic models as narratives, namely being able to infer how a course of events, decomposed into steps, may change if we change one step in particular. Mental animation is a process of complex visual-spatial inference. Limits and capabilities of humans in such tasks depend on the cognitive architecture of human mind[5]. What Hegarty has found is that mental animation is *piecemeal*, in the sense that human mind does not animate the components of a diagram in parallel, but rather infer the motion of components *one by one*. This strategy has a straightforward consequence; in order to proceed with animating components, we should store intermediate results of inferences drawn on previous components. Due to the limitations of working memory (WM), people usually store such information on external displays. Hegarty has provided evidence that diagrammatic reasoning is bounded to WM abilities. The more we proceed in inferring animation on later components, the more the inferences on earlier components degrade (see for instance Figure 2); "as more components of the system are 'read into' spatial working memory, the activation of all items is degraded, so that when later components are in, there is not enough activation of the later components to infer their motion" (Hegarty 2000, p 201).

[5] On this, I rely on the framework assumed by the cognitive-load theory (Paas et al 2010)

a



b

Cycle	Action Taken
1	Rope Strand 1 is inspected (simulation is given knowledge that this rope strand is being pulled)
2	Rope Strand 1 is animated (i.e. its direction of motion is inferred)
3	Rope Strand 2 is inspected
4	Rope Strand 2 is animated
5	Rope Strand 3 is inspected Pulley 1 is inspecte
6	Rope Strand 3 is animated Pulley 1 is animated
7	Rope Strand 4 is inspected
8	Rope Strand 4 is animated
9	Rope Strand 5 is inspected Pulley 2 is inspected
10	Rotation of Pulley 2 is inferred Translation of Pulley 2 is inferred
11	Weight is inspected
12	Weight is animated

Figure 2. (a) Example of diagram of a simple pulley system that can be mentally animated (b) Description of typical actions that can be one by one to animate the pulley system. Both figures taken from Hergaty (2000)

The actual limit of our cognitive architecture on this respect may be debated, and it is an empirical issue. The important point is that *no matter our external displays*, for very large systems (such as Figure 3) it is very unlikely that human cognition will be able to process all information about elements interactivity. This is because by animating components one-by-one, even if we use sophisticated instruments such computer simulations, still inferences on earlier components will degrade. This means that build-it tests will be very ineffective, if not impossible. In terms of narratives, recipes and mechanistic models, this means that for large mechanistic diagrams with many model components, no human would be able to anticipate the consequences of modifying a step in the model for all the other steps of the model, even if a computer simulation shows that the phenomenon can be possibly produced by the complex model. The computer simulation may highlight certain aspects (as Bechtel in 2016 notes), but the model is not intelligible in the sense required by mechanistic philosophy. *If the model is not intelligible in this way, then it cannot be possibly turned into an explanation*.

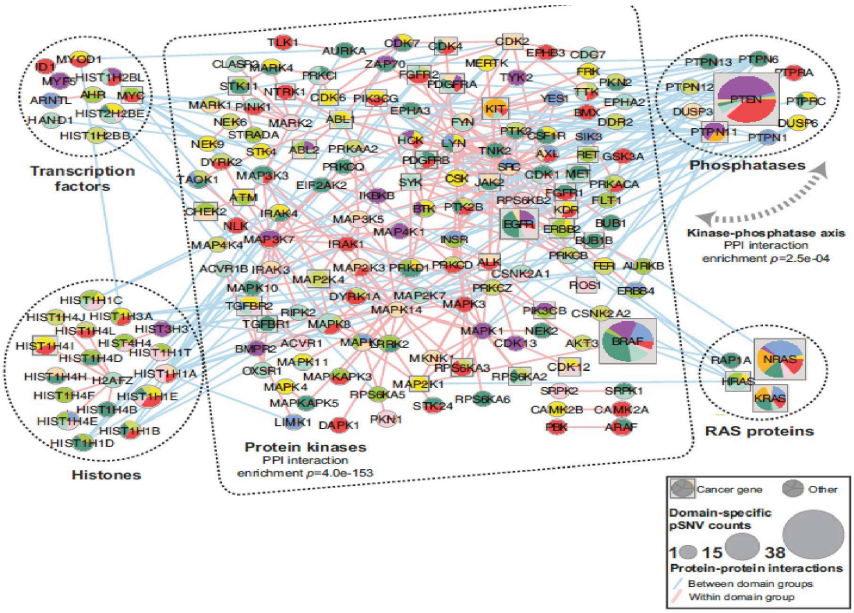


Figure 3. Network of interactions of proteins with significant enrichment of phosphorylation-related single nucleotide variations. Phosphorylation is a central post-translational modification in cancer biology. Authors are not trying to recompose the mechanism that from phosphorylated proteins (nodes) lead to a tumor phenotype, but rather to identify the magnitude of the impact of this process on cancer genes. Figure taken from (Reimand et al 2013)

The results of Hegarty's research suggest that when mechanistic models are concerned, strategies of localization are effective (in terms of explanatory potential) only when a limited number of model components are actually identified. The number may increase if we use computer simulations. However, for very large amounts of model components (such as Figure 3) recomposition is just impossible for humans, because inferences on the role of components in the causal division of labor of a phenomenon will degrade to make place for inferences about other components. This of course holds only if we have explanatory aims in mind.

To summarize, in section 2 we have made three claims:

1. If a how-possibly model can be turned into an explanation, then it is intelligible
2. If a model is not intelligible, then it cannot become explanatory
3. Complex models are a class of non-intelligible models

3. MACHINE LEARNING AND LOCALIZATIONS

Machine learning (ML) is a subfield of computer science which studies the design of computing machinery that improves its performance as it learns from its environment. A ML algorithm extracts knowledge from the input data, so that it can give better solutions to the problem that it is meant to solve. This learning process usually involves the automatic construction and refinement of a model of the incoming data. In ML terminology, a model is an information structure which is stored in the computer memory and manipulated by the algorithm.

As mentioned before, the concept of 'problem' in ML has a specific meaning which is different from other fields of science. A ML problem is defined by a set of input variables, a set of output variables, and a collection of samples which are input-output pairs. Solving a problem here means finding a quantitative relation between inputs and outputs in the form of a predictive model, in the sense that the algorithm will be used to produce a certain output given the presence of a specific input.

3.1 The PARADIGM algorithm

ML has been applied in the molecular sciences in many ways (Libbrecht and Noble 2015). Especially in cancer research[6], computer scientists have created and trained a great deal of algorithms in order to identify entities that are likely to be involved in the development of tumors, how they interact, to predict phenotypes, to recognize crucial sequences, etc (see for instance Leung et al 2016).

As a topical example of ML applied to biology, we introduce an algorithm called PARADIGM (Vaske et al 2010). This algorithm is used to infer how genetic changes in a patient influence or disrupt important genetic pathways underlying cancer progression. This is important because there is empirical evidence that "when patients harbor genomic alterations or aberrant expression in different genes, these genes often participate in a common pathway" (Vaske et al 2010, p i237). Because pathways are so large and biologists cannot hold in their mind the entities participating in them, PARADIGM integrate several genomic datasets – including datasets about interactions between genes and phenotypic consequences – to infer molecular pathways altered in patients; it predicts

[6] See for instance The Cancer Genome Atlas at https://cancergenome.nih.gov

whether a patient will have specific pathways disrupted given his/her genetic mutations.

The algorithm is based on a simplified model of the cell. Each biological pathway is modeled by a graph. Each graph contains a set of nodes, such that each node represents a cell entity, like a mRNA, a gene or a complex. A node can be only in three states (i.e. activated, normal or deactivated). The connections among nodes are called factors, and they represent the influence of some entities on other entities. It must be noticed that the model does not represent why or how these influences are exerted. Only the sign of the influence, i.e. positive or negative, is specified.

The model specifies how the expected state of an entity must be estimated. The entities which are connected by positive or negative factors to the entity at hand cast votes which are computed by multiplying +1 or –1 by the states of those entities, respectively. In addition to this, there are 'maximum' and 'minimum' connections to cast votes which are the maximum or the minimum of the states of the connected entities, respectively. Overall, the expected state of an entity is computed as the result of combining several votes obtained from the entities which are connected to it. Such a voting procedure can be associated to localizations (i.e. whether a node is activated or not), but hardly to biological explanations.

The states of the entities can be hidden, i.e. they can not be directly measured on the patients, or observable. The states of the hidden variables must be estimated by a probabilistic inference algorithm, which takes into account the states of the observed variables and the factors to estimate the most likely values of the hidden variables. Here it must be pointed out that this algorithm does not yield any explanation about the computed estimation. Moreover, it could be the case that the estimated values are not the most likely ones, since the algorithm does not guarantee that it finds the globally optimum solution.

The size of the model is determined by the number of entities and factors that the scientist wishes to insert. A larger model provides a perspective of the cell processes which contains more elements, and it might yield better predictions. This means that the more components the model has, the better the algorithm will perform. In biological terms, the larger the model, the more precise *complex localizations* the algorithm will identify, in particular by pointing more precisely towards pathways that are likely to be

disrupted in the patient with more information about the state of gene activities, complexes and cellular processes. Importantly, PARADIGM does not infer new genetic interactions, but it just helps identifying those known interaction in a new data set. It is completely supervised, in the sense that "[w]hile it infers hidden quantities (…), it makes no attempt to infer new interactions not already present in an NCI [National Cancer institute database] pathway" (Vaske et al 2010, p i244).

4 COMPLEX MODELS AND MECHANISTIC EXPLANATIONS

Before unwinding our conclusions, let me recall the results of Section 2 very briefly:

1. If a how-possibly model can be turned into an explanation, then it is intelligible[7].
2. If a model is not intelligible, then it cannot become explanatory
3. Complex model (in the sense explained in 2.2) are not intelligible

What does this have to do with PARADIGM? It is important to emphasize what we have pointed out in Section 3.1, namely that an algorithm like PARADIGM is more efficient when working with more components. If we think about models generated by algorithms such as PARADIGM in mechanistic terms, this means that the algorithm provides more precise complex localizations, because more entities that are likely to be causally relevant to a phenomenon are identified, and the information about the probability of a pathway being disrupted in a patient will be more precise. However, the models will be more complex, and they will be decreasingly intelligible. This is because the final model will count an elevated number of components, and recomposing these components into a full-fledged mechanistic explanation of how a tumor is behaving will be cognitively very difficult; the inferences about the behavior of components are not run in parallel, but one by one, and once we proceed in inferring the behavior of a component on the basis of the behavior of another component, other inferences will degrade, as Hegarty's studies have shown. In the ideal situation, PARADIGM will generate unintelligible models:

[7] Remember: A mechanistic model *x* is intelligible to a modeler *y* if *y* can use the information about the components of *x* to instantiate so-called 'build-it test'. Such tests are performed on how-possibly models to turn them into explanatory models by obtaining information on how to recompose a phenomenon (i.e. by showing how a list of biological entities are organized to produce a phenomenon).

4. Algorithms such as PARADIGM generate models which are not intelligible because such models are too complex
5. Because of 2, 3 and 4, complex models generated out of algorithms like PARADIGM cannot become explanations

This means that when we use algorithms such as PARADIGM to cope with the complexity of biological systems, we successfully handle big data sets, but such a mastery comes at a price. Using ML in molecular biology means providing more detailed localizations, but we also lose explanatory power, because no modeler will be able to recompose the mechanism out of a long list of entities.

This implies that, in the mechanistic epistemic horizon, the central role assigned to explanations should be reconsidered when contemporary molecular biosciences are concerned. As Bechtel has also emphasized in the context of computational models in mechanistic research (2016), such tools are useful to show whether some entities are likely to be involved in a particular phenomenon or suggest alternative hypotheses about the relation between certain entities. However, providing fully-fledged mechanistic explanations is another thing. It is the same with algorithms of ML; we identify more entities likely to be involved in a mechanism, we may even find out that entities involved in specific process may be connected with entities involved in other processes (via for instance Gene Ontology enrichments), but we cannot recompose a mechanism out of a list of hundreds of entities. In fact, we come to value different epistemic values, and *explanatory power is not one of them*. This somehow implies also a shift in the way scientific articles are organized; if in 'traditional' molecular biology evidence converges towards the characterization of a single mechanism, in data-intensive biology we make a list of entities that can be involved in a phenomenon, but we do not necessarily connect those entities mechanistically (Alberts 2012). Another strategy (Krogan et al 2015) – though motivated more by biologically rather than cognitive reasons – is to abstract from macromolecular entities and consider only aggregates of them in the form of networks; whether establishing network topology is providing a mechanistic explanation remains an open question.

REFERENCES

Alberts, B. (2012). The End of "Small Science"? *Science*, *337*(September), 1230529.

Bechtel, W. (2016). Using computational models to discover and understand mechanisms. *Studies in History and Philosophy of Science Part A*, *56*, 113–121.

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity - Decomposition and Localization as Strategies in Scientific Research*. Cambridge, Massachusetts, and London, England: The MIT Press.

Craver, C. (2007). *Explaining the Brain - Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. Chicago: The University of Chicago Press.

De Regt, H. (2017). *Understanding Scientific Understanding*. Oxford: Oxford University Press.

de Regt, H. W. (2015). Scientific understanding: truth or dare? *Synthese*, *192*(12), 3781–3797. http://doi.org/10.1007/s11229-014-0538-7

Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese*, *172*(2), 269–281.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford University Press.

Hegarty, M. (2000). Capacity Limits in Mechanical Reasoning. In M. Anderson, P. Cheng, & V. Haarslev (Eds.), *Diagrams 2000* (pp. 194–206). Springer-Verlag.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A., & Ideker, T. (2015). The Cancer Cell Map Initiative: Defining the Hallmark Networks of Cancer. *Molecular Cell*, *58*(4), 690–698.

Levy, A. (2014). What was Hodgkin and Huxley's achievement? *British Journal for the Philosophy of Science*, *65*(3), 469–492.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about Mechanisms. *Philosophy of Science*, (67), 1–25.

Morrison, M., & Morgan, M. (1999). Models as mediating instruments. In M. Morrison & M. Morgan (Eds.), *Models as Mediators*. Cambridge University Press.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: The MIT Press.

Ratti, E. (2018). "Models of" and "models for": On the relation between mechanistic models and experimental strategies in molecular biology. *British Journal for the Philosophy of Science*.

Reimand, J., Wagih, O., & Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Scientific Reports*, *3*.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., … Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, *26*(12), 237–245.

Vasu, K., & Nagaraja, V. (2013). Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiology and Molecular Biology Reviews*, *77*(1), 53–72.

The Roles of Possibility and Mechanism in Narrative Explanation

Abstract

There is a fairly longstanding distinction between what are called the *ideographic* as opposed to *nomothetic* sciences. The nomothetic sciences, such as physics, offer explanations in terms of the laws and regular operations of nature. The ideographic sciences, such as natural history (or, more controversially, evolutionary biology), cast explanations in terms of narratives. This paper offers an account of what is involved in offering an explanatory narrative in the historical (ideographic) sciences. I argue that narrative explanations involve two chief components: a possibility space and an explanatory causal mechanism. The presence of a possibility space is a consequence of the fact that the presently available evidence underdetermines the true historical sequence from an epistemic perspective. But the addition of an explanatory causal mechanism gives us a reason to favor one causal history over another; that is, causal mechanisms enhance our epistemic position in the face of widespread underdetermination. This is in contrast to some recent work that has argued against the use of mechanisms in some narrative contexts. Indeed, I argue that an adequate causal mechanism is always involved in narrative explanation, or else we do not have an explanation at all.

1. Introduction

The historical sciences (geology, paleontology, evolutionary biology, etc.)[1] are usually thought to deploy different explanatory strategies than the non-historical sciences (Turner 2007; Turner 2013). Whereas physics, say, seeks explanations given in terms of general laws and the like, the historical sciences seek to explain in terms of narratives. In this paper I will argue for a version of narrative explanation involving two chief components: possibility spaces and causal mechanisms. It has recently been argued that complex historical narratives (to be defined later) can't support explanations involving causal mechanisms (Currie 2014). I argue that this is mistaken. I'll go over some recent work on the history of abiogenesis research to support this contention.

The argument presented in this paper will defend two primary claims: (1) the conceptual structure of narrative explanations nearly always involves a space of alternative possibilities. This can be for either epistemic or ontological reasons. From and epistemic perspective possibility spaces are necessary on account of our position relative to the available evidence. That is, the available evidence radically underdetermines any particular causal history, and on the basis of that fact many possible histories appear compatible with what we know (see Gordon and Olson 1994, p. 15). Construed ontologically, a set of historical facts might involve a high degree of objective contingency—it might be the case that things really could have gone a number of different ways. For the purposes of this paper I remain silent with respect to this ontological aspect and defend the importance of possibility spaces for largely epistemic reasons. (2) Adequate causal mechanisms enhance our epistemic position relative to alternative causal

[1] I note that the idea of evolutionary biology as a properly "historical science" is a controversial one. See Ereshefsky (1992) for some strong arguments against the idea of evolutionary biology as having a distinctively 'historical' flavor.

histories. Causal mechanisms put us in a position to better assess the plausibility of a given history within our possibility space, and in this way enhance the epistemic power of a purportedly explanatory historical narrative. This can involve either the actual discovery of such mechanisms, or raw theoretical innovation. Citing an adequate causal mechanism may not discriminate between possibilities in decisive fashion. Rampant underdetermination seems to rule out such a possibility (see Turner 2007). But an adequate mechanism does make a given explanation more explanatory than its competitors, and so part of the task is to see how this notion of mechanistic adequacy can be cashed out in such a way as to make this notion of *explanatoriness* epistemically significant and not simply *ad hoc*.

2. The Role of Possibility Spaces

In the introduction I said that I would defend two major claims: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. This section will address the first claim by giving a more detailed account of the conceptual structure of narrative explanations and why the role of possibility spaces is so central to them.

When confronted with a natural historical problem (e.g. accounting for the processes involved in the formation of atoll reefs, say (see Ghiselin 1969)) it is my claim that what we are confronted with is, in fact, a space of *possible* histories. That is, when the historical scientist attempts to answer the question, "What geological process accounts for the formation of atoll reefs?" she understands—perhaps implicitly—that there are a number of ways things *might* have gone: she sees many possible histories. This space of possible histories essentially generates a contrasting set of possible explanations, each possible history corresponding roughly to one

hypothetical solution to the problem.[2] Obviously there's just one causal history that actually obtained, but the evidential situation is such that this history is not uniquely fixed from an epistemic perspective (see Roth 2017). The historical scientist's explanatory task then consists in finding the best approximation of the true causal history.

A nice example of this sort of reasoning process can be glimpsed in the debates over speciation processes among evolutionary biologists and paleobiologists. Stephen J. Gould and Niles Eldredge (1972) developed the theory of *punctuated equilibria* to account for the pattern of speciation witnessed in the fossil record. The idea of punctuated equilibria, in brief, holds that evolutionary change occurs in sudden bursts (on geological timescales, anyway), followed by long periods of relative evolutionary stasis. The going theory of evolutionary change at the time held to *phyletic gradualism*—the idea that the pace of evolution is slow and relatively uniform (see Turner 2011). Each of these alternatives is broadly consistent with the available fossil evidence. Phyletic gradualism takes the view that the evolutionary process is gradual, and that the fossil record is very patchy. The putatively patchy character of the fossil record means that we shouldn't expect to be able to use it as a tool for faithfully reading off patterns of speciation in the actual history of life. The theory of punctuated equilibria has it that the fossil record is relatively faithful to evolutionary history, meaning that the fossil record *does* have some explanatory import with respect to uncovering important evolutionary patterns (like speciation). The evidence in the fossil record can support either interpretation.

Consider another example, this time from geology. 19th century geologists were confronted with a fascinating geological puzzle involving what were called 'erratic blocks'.

[2] I'm certainly *not* claiming that the historical scientist is in a position to generate or realize all possible histories, as the number of such alternatives is plausibly infinite. But certainly it's possible to generate quite a few, and it seems that in fact we usually do.

These hulking slabs of (usually) granite are found miles away from any related rocks, and so the obvious question to be answered is, "How did such a large piece of granite come to be deposited here?" In 1820s Europe the answer was not immediately obvious. One well-documented case involved a granite erratic in Switzerland, which was determined to be composed of primary rocks of Alpine origin, but resting on a limestone formation many hundreds of miles from any mountains (see Rudwick 2014, pp. 117-25). Several explanations were offered: that it was deposited by the waters of the Noahic deluge; that it was carried and deposited by waters traveling down the Alps from a broken mountain dam; and only later that it was carried by glacial ice and then deposited after a subsequent melt. The process of adjudicating between each such purportedly explanatory histories (whether evolutionary patterns or seemingly bizarre geological deposits) is the subject of the next section.

It's important to stress that the evidential underdetermination of historical hypotheses is quite different than underdetermination in science more broadly. Turner (2007) argues convincingly that the problem of underdetermination is rather severe in the historical sciences given that natural processes actively destroy the evidential traces on which historical scientists rely.[3] There are two points that make this worthy of note. First, it is precisely for this reason that the explanatory task of the historical scientist *necessarily* involves the generation of a possibility space. If we can think of a natural history as a story concerning the artifacts of the natural world, then what the world presents us with is a story that's missing a great many pages. The unfortunate fact of the matter is that there are many ways of filling those pages in, each of which

[3] Turner appeals to the role played by background theories in the historical sciences to motivate his point. Here, there relevant theory is *taphonomy*, which describes the mechanisms by which the relevant evidence is destroyed (remineralization, decomposition, etc.).

is broadly compatible with our evidential situation.[4] Second, widespread underdetermination is what motivates the earlier insight that the explanatory aspiration of historical science is to give the best *approximation* of the true causal history. It is implausible to think that any of the historical hypotheses we generate will fill in the missing pages perfectly, but we can have reasons to think that some hypotheses outperform others (of which more to come).

To summarize, possibility spaces are ineliminable from narrative explanations because of our epistemic position relative to the evidence at hand. What we want is to develop a causal history that explains the phenomenon in question (e.g. erratic blocks and evolutionary patterns), but right away we realize that many different and mutually incompatible histories could—hypothetically—do the trick. The construction of a space of live possibilities allows us to have some degree of confidence that we've explored the relevant alternatives.[5] Once we've developed a space of possibilities, the initial question (such as, "What accounts for the formation of atoll reefs?") becomes importantly *contrastive*: "Why *x* and not *x'*?" where *x* and *x'* are alternative possible causal histories accounting for the target phenomenon. We want to know how it is that possibilities come to be "foreclosed" upon as a narrative explanation develops, as Beatty (2016) puts it.

3. Causal Mechanisms and Hypothesis Adjudication

[4] See Turner (2011) chapter 2 for more in-depth discussion.

[5] There's a way of reading this that might tempt one to see this as something akin to *inference to the best explanation*. Any such connection is largely superficial. The primary reason for this is that the explanatory scheme that I'm outlining is not meant to be making any especially strong claims about the strength of an explanation as related to its connection to reality. Perhaps none of the causal histories we generate are very accurate as descriptions of the true causal history.

7

I now turn my attention to an explication and defense of (2): adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. Causal mechanisms are what provide reasons for preferring one possible causal history over another as regards the space of possible histories generated by the natural historical problem at hand.

3.1. Mechanistic set-ups-

Because contingency is generally seen as playing such a fundamental role in natural historical contexts, the relevant mechanisms are not likely to be cashed out in terms of 'invariances' and 'regularities,' as is common in other scientific contexts (see Havstad 2011; Darden and Craver 2002). For the purposes of natural history we might instead think in terms of a more minimal conception of causal mechanisms that I'll call *mechanistic set-ups*. A mechanistic set-up differs from paradigmatic mechanisms (as in Glennan (2002))[6] in that it will often be the case that mechanistic set-ups are the result of one-off circumstances. Paradigmatic mechanisms characterize causal systems that are largely stable across time (think of protein synthesis, for instance). Mechanistic set-ups are not stable across time in this way, but still render outcomes causally expectable given that the right antecedent conditions obtain. That is, given that the right antecedent conditions obtain (and this may, of course, be a *highly contingent* affair), the causal output of the system is fully determined—we have a case of mechanical causal output.

Nancy Cartwright and John Pemberton (2013) give a simple example of a mechanistic set-up using a toy sailboat. When the toy boat is placed in the water it displaces enough liquid to

[6] "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations."

stay afloat; it has a wind-catching device for locomotion; the wind-catching device is acted about by wind gusts in order to achieve locomotive action. If we take this example as having to do with the actions of an *agent* that brings about the mechanistic set-up then we might incline toward an interpretation of the situation in terms of paradigmatic mechanisms. But imagine there's no agent involved at all; that is, let it be the case that nobody placed the boat on the water, and likewise nobody chose any windy day in particular for the use of the boat. Instead suppose that it is a series of contingent events (a child threw the boat in the garbage, it fell out of the garbage truck on the highway, and is now on the surface of a local pond, etc.) that have made things such that the boat is at some later time moving across the top of the water in the expected way.

The one-offness of the circumstances in the revised toy boat example doesn't seem to make the situation non-mechanistic in character. Rather, the mechanism just isn't stable across time in the same way paradigmatic mechanisms are. This is a mechanism in a more minimal sense: it is a mechanistic set-up. In other words, the realization of appropriate antecedent conditions renders the outcome causally expectable, even though the antecedent conditions are highly contingent.[7]

This case is so simple that it won't have much bite against Currie. Recall that Currie's claim is that mechanisms show to be of no use in *complex* narratives. In these cases the explanatory targets are *diffuse*, meaning that they involve complex networks of causal contributors (Currie 2014). An example of a diffuse target is Sauropod gigantism, Gigantism involves, at least, skeletal pneumatization, ovipary, increased basal metabolic rate, etc. Nothing seems to unify such causal contributions, and so there is no *mechanism* for gigantism, according to Currie—the explanatory target is *too diffuse* in complex narratives.

[7] See chapter 3 of Conway Morris (2003) for an in-depth discussion.

3.2. Abiogenesis, mechanistic set-ups, and hypothesis adjudication-

Abiogenesis, I argue, qualifies as a minimal mechanistic set-up in the sense just argued for. That is, the set of facts that determined the development of the very first self-replicating, heterotrophic organisms are plausibly subject to a high degree of contingency (see Conway Morris 2003), but even so, life is a deterministic consequence of just such a contingent set of facts.[8] Further, the instances that the theory aims to explain (e.g. self-replicating molecular systems; heterotrophic metabolic systems; protective membrane enclosures, etc.) are diffuse in the same sense as Sauropod gigantism. My aim here is not to give a full theoretical survey of abiogenesis, but instead to provide just enough content to justify the claim that work in this area fulfills the description of narrative already given, and that causal mechanisms play an important explanatory role, specifically to do with hypothesis adjudication.

Probably the first serious theoretical work on the origins of life is A.I. Oparin's 1923 *The Origins of Life* (Falk and Lazcano 2012). The basic theoretical framework is familiarly Darwinian. Oparin had in mind a model of biological origins whereby life comes on-line in stages, rather than all at once. The prebiotic world, on this view, was one of something approximating 'molecular competition.' For Oparin this amounted to chemical assemblages witnessing differential stability, approximately underwriting a growth model of molecular evolution (Falk and Lazcano 2012; Pigliucci 1999). The primary thing to be explained, on this model, was the development of heterotrophic metabolism. Metabolic pathways are so complex

[8] Some recent work in origins of life research may end up giving reasons to question the assumed contingency of life's emergence. See Kauffman (1993) for a classic treatment of the "self-organization" thesis, and England (2015) for more recent theoretical developments.

that Oparin thought their development must be accounted for in a basically stepwise fashion. Differential stabilities of chemical assemblages would make it such that certain molecules would make up increasingly large proportions of the chemical 'population,' making them live candidates for further downstream innovation (like complex metabolic pathways).

Oparin-type selection models have mostly—though perhaps not entirely—fallen by the wayside. Contemporary work is focused primarily on accounting for the possibility of self-replication and autocatalysis (Penny 2005). The thought is that biological origins must be accounted for in something like a two-step process, one involving the development of self-replicating material suitable for hereditary mechanisms, and another for things like metabolism and heterocatalytic functions like protein construction (Falk and Lazcano 2012; Conway Morris 2003). One of the more promising research strains in this area concerns what's known as the 'RNA World' (Conway Morris 2003). It's widely believed to be the case that the first replicators were RNA (or RNA-like) molecules. So, RNA World researchers are attempting to simulate the conditions of the prebiotic Earth in the laboratory in order to see whether the RNA model of biological origins can carry its empirical weight.

Of note for the purposes of this paper is that the dispute between metabolism-first and replication-first models of abiogenesis is precisely over whether the causal mechanisms in play can adequately account for the target phenomenon: namely, the development of living organisms in the ancient history of Earth. H.J. Muller developed a theoretical agenda stressing the need for self-replicators at the historical foundations of life (Falk and Lazcano 2012). Oparin took heterotrophic metabolic pathways as the primary puzzle to be solved (Oparin 1938; Falk and Lazcano 2012). The replication-first view has emerged as the going view among contemporary researchers primarily because it offers a more plausible mechanism for life's early development.

In order to build complex metabolic pathway it seems like it's first necessary to have a genome space that's large enough to enable downstream innovation of complex functions. So it is that the replication-first view and the research agenda dictated by projects like RNA World are taken to be more explanatory than Oparin-type explanations given in terms of selection among molecular assemblages.

4. Putting Things Together

Let's recall once more the two key claims being advanced: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. Widespread underdetermination in the historical sciences leads to the persistent appearance of possibility spaces as specified by (1), and the development of adequate causal mechanisms specified under (2) enhances our ability to adjudicate the alternatives we're faced with. Causal mechanisms put us in a position to address the contrastive question, "Why *x* and not *x'*?" Causal mechanisms are the devices by which historical counterfactuals become foreclosed upon in the sense of Beatty (2016).

Because explanation in the historical sciences is contrastive in the above sense, I argue that some notion of mechanism is involved in *every* case of successful narrative explanation. Currie (2014) argues that causal mechanisms are appropriate only for the purposes of simple narratives apt to be embedded in terms of regularities. Complex narratives with their diffuse explanatory targets require something more piecemeal that doesn't count as a causal mechanism. My more minimal conception of causal mechanisms given in terms of *mechanistic set-ups* sheds light on why this can't be right. Mechanistic set-ups aren't stable across time like paradigmatic

mechanisms, and yet we have good reason to think that the consequences of such set-ups are mechanistically determined (see Penny 2005; Glennan 2010).[9] It is just this sort of conception of mechanism that helps us to make sense of explanatory success in abiogenesis (such as it is).

Surely the genesis of the first biotic creatures is every bit as diffuse an explanatory target as Sauropod gigantism. I've argued (and I think convincingly) that it is precisely due to the adequacy of some underlying mechanism that one explanatory agenda in abiogenesis has been accepted over the alternatives. The complexity of the narrative and the diffuseness of the explanatory target appear to be beside the point. Without an adequate mechanism—however minimally construed—we can't answer the contrastive question, and so we have no explanation at all.

5. Objection and a Reply

According to Currie (2014) mechanistic set-ups (*ephemeral mechanisms* (Glennan 2010)) look like they're simply pointing to claims about sensitivity to initial conditions. If that's right, then there's a problem, because causal processes in natural historical contexts are often thought to be contingent not just in the sense that they display sensitivity to initial conditions. Such processes are taken to be subject to contingencies in a more robust sense involving "causal cascades" themselves (Currie 2014). It is not unreasonable, for instance, to think that whether a chemical assemblage will manage to hit the right configuration and produce a self-replicating RNA strand is not just a matter of realizing the right set-up conditions (independent of the chances of hitting

[9] Penny notes some interesting experimental results in which living organisms are frozen to near absolute zero, meaning that all information concerning the positions and velocities of the particles in their make-up is lost. They can, nonetheless, be successfully reanimated. Given that the only information that's retained after such a deep freezing involves the chemical structure of the organisms, a natural inference is that 'life' is a mechanical consequence of chemical parts.

on such a configuration). Whether the chemical elements enter into the appropriate causal relations for manifesting autocatalysis might *itself* be a probabilistic matter. Having the right elements might not be all you need—you might need the right elements plus a bit of probabilistic luck. Objective probabilities of this sort might do some damage to the mechanistic account, since it would seem not to be the case that an explanandum *just follows* from a causal set-up. The force of this objection is at least partly dependent on one's answer to the question of where in the world we ought to 'place' objective chances (if there are any).

Most of our intuitions about objective probabilities (probably) derive from our ongoing observations of the world. A lot of stuff in the world *just seems* chancy. We regularly speak in terms of the "odds" or "chances" of developing cancer and the like. Simplifying quite a bit, when we say that there's a 40 percent chance that Susan will live for more than 5 years after being diagnosed with some cancer that has developed to some particular stage, what we're saying is that approximately 40 percent of people that present as cases sufficiently similar to Susan have lived for 5 years or more. One way to read this is in terms of causal indeterminacy. That is, there is really no matter of the fact at time *t* as to what will be the case at time *t'*, aside from the probabilistic facts about cancer populations. The future is (to some degree) causally open, as the causal cascades are operating in a fundamentally probabilistic way.

Such a reading, however, is by no means forced. Bruce Glymour (1998) offers a picture wherein objective probabilities are placed at the level of causal *interactions*. That is, entities *e* and *e** enter into causal interactions with each other on a probabilistic basis, but when they do, the downstream effects unfold in a fully deterministic fashion. Probabilistic partitions of the world, then, are just reflections of whether certain causal interactions became manifest in certain subpopulations or not. If 40 percent of patients with a certain cancer at a particular stage will

survive for more than five year, it's because free radicals (probabilistically) failed to enter into certain causal interactions with healthy cells. The opposite is the case for the contrasting class of fatal cases. On this picture, determinism of the relevant kind seems to be preserved. In such cases as the right causal interactions are realized, downstream effects unfold in mechanical fashion.

6. Conclusion

In this paper I argued for two main claims: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. The reason that narrative explanations involve possibility spaces has to do with our epistemic position relative to the available evidence. Undetermination so permeates the historical sciences that any problem for which we seek an explanation will involve an array of possible alternative causal histories, each of which is broadly consistent with the available evidence. It is the introduction of an adequate causal mechanism that puts us in a position to improve our epistemic lot—with a good mechanism in hand, we can begin to foreclose upon alternatives.

References

Beatty, John. 2016. “What Are Narratives Good For?” *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 58. Elsevier Ltd: 33–40. doi:10.1016/j.shpsc.2015.12.016.

———. 2017. “Narrative Possibility and Narrative Explanation.” *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 1–14. doi:10.1016/j.shpsa.2017.03.001.

Cartwright, Nancy, and John Pemberton. 2013. "Aristotelian Powers: Without Them, What Would Science Do?" in Groff & Greco (Eds.), *Powers and Capacities in Philosophy: The New Aristotelianism*. New York: Routledge.

Conway Morris, Simon. 2003. *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge: Cambridge University Press.

Currie, Adrian Mitchell. 2014. “Narratives, Mechanisms and Progress in Historical Science.” *Synthese* 191 (6): 1163–83. doi:10.1007/s11229-013-0317-x.

Darden, Lindley, and Carl Craver. 2002. "Strategies in the interfield discovery of the mechanism of protein synthesis." *Studies in History and Philosophy of Biological and Biomedical Sciences* 33: 1-28.

Eldredge, Niles, and Stephen J. Gould. 1972. "Punctuated equilibria: an alternative to phyletic gradualism," in Schopf (Ed.), *Models in Paleobiology*. San Francisco: Freeman Cooper.

England, Jeremy. 2015. "Dissipative Adaptation in Self-Driven Assembly." *Nature Nanotechnology*, 10: 919-923.

Ereshefsky, Marc. 1992. "The Historical Nature of Evolutionary Theory." In *History and Evolution*, ed. Matthew Nitecki and Doris Nitecki. New York: The SUNY Press.

Falk, Raphael, and Antonio Lazcano. 2012. “The Forgotten Dispute: A.I. Oparin and H.J. Muller on the Origin of Life.” *History and Philosophy of the Life Sciences* 34 (3): 373–90.

Ghiselin, Michael T. 1969. *The Triumph of the Darwinian Method*. Chicago: Chicago University Press.

Glennan, Stuart. 1996. “Mechanisms and the Nature of Causation.” *Erkenntnis* 44 (1): 49–71.

———. 2002. “Rethinking Mechanistic Explanation.” *Philosophy of Science* 69 (S3): S342–53.

———. 2010. “Ephemeral Mechanisms and Historical Explanation.” *Erkenntnis* 72 (2): 251–66. doi:10.1007/s10670-009-9203-9.

Glymour, Bruce. 1998. “Contrastive, Non-Probabilistic Statistical Explanations.” *Philosophy of Science* 65 (3): 448–71.

Gordon, Malcolm and Everett Olson. 1994. *Invasions of the Land*. New York: Columbia University Press.

Haldane, J.B.S. 1954. "The origin of life." *New Biology* 16: 12-27.

Havstad, Joyce C. 2011. “Problems for Natural Selection as a Mechanism.” *Philosophy of Science* 78 (3): 512–23. doi:10.1086/660734.

Hull, David. 1975. “Central Subjects and Historical Narratives.” *History and Theory* 14 (3): 253–74.

Jeffares, Ben. 2008. “Testing Times: Regularities in the Historical Sciences.” *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 39 (4). Elsevier Ltd: 469–75. doi:10.1016/j.shpsc.2008.09.003.

Kauffman, Stewart. 1993. *The Origins of Order: Self Organization and Selection in Evolution*. Oxford: Oxford University Press.

Mink, Louis O. 1970. "History and Fiction as Modes of Comprehension." *New Literary History*, 1 (3): 541-558.

Oparin, A.I. 1938. *The Origin of Life*. New York: MacMillan.

Penny, David. 2005. “An Interpretive Review of the Origin of Life Research.” *Biology & Philosophy* 20 (4): 633–71. doi:10.1007/s10539-004-7342-6.

Pigliucci, Massimo. 1999. "Where do we come from? A humbling look at the biology of life's origin." *Skeptical Inquirer* 99: 193-206.

Ricoeur, Paul. 1984. *Time and Narrative (Volume 1)*. Chicago: University of Chicago Press.

Roth, Paul A. 2017. “Essentially Narrative Explanations.” *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 1–9. doi:10.1016/j.shpsa.2017.03.008.

Rudwick, M.J.S. 2014. *Earth's Deep History: How It Was Discovered and Why It Matters*. Chicago: Chicago University Press.

Sepkosi, David. 2012. *Rereading the Fossil Record: The Growth of Paleontology as an Evolutionary Discipline*. Chicago: Chicago University Press.

Sunstein, Cass R. 2016. “Historical Explanations Always Involve Counterfactual History.” *Journal of the Philosophy of History* 10 (3): 433–40. doi:10.1163/18722636-12341345.

Turner, Derek. 2013. “Historical Geology: Methodology and Metaphysics.” *Geological Society of America Special Papers* 502 (2): 11–18. doi:10.1130/2013.2502(02).

———. 2007. *Making Prehistory: Historical Science and the Scientific Realism Debate*. Cambridge: Cambridge University Press.

———. 2011. *Paleontology: A Philosophical Introduction*. Cambridge: Cambridge University Press.

18

A Better Foundation for Public Trust in Science

S. Andrew Schroeder
Claremont McKenna College/Princeton University
aschroeder@cmc.edu

draft of 15 June 2018

Abstract. There is a growing consensus among philosophers of science that core parts of the scientific process involve non-epistemic values. This undermines the traditional foundation for public trust in science. In this paper I consider two proposals for justifying public trust in value-laden science. According to the first, scientists can promote trust by being transparent about their value choices. On the second, trust requires that the values of a scientist align with the values of an individual member of the public. I argue that neither of these proposals work and suggest an alternative that does better: when scientists must appeal to values in the course of their research, they should appeal to *democratic values*, the values of the public or its representatives.

1. Introduction

The American public's trust in science is a complicated matter. Surveys reveal that trust in science has remained consistently high for decades, and scientists remain among the most highly-trusted professional groups (Funk 2017). However, within some segments of society (especially conservatives) trust has declined significantly (Gauchat 2012), and there are obviously serious gaps in trust on certain issues, such as climate change, vaccine safety, and GM foods (Funk 2017). The picture, then, is a complex one, but on balance it is clear that things would be better if the public placed greater trust in science and scientists, at least on certain issues.

As a philosopher, I am not in a position to determine what explains the lack of trust in science, nor to weigh on what will in fact increase trust. Instead, in this paper I will look at the question of what scientists can do to *merit* the public's trust — under what conditions the public *should* trust scientists. Indeed, it seems to me that we need to answer the normative question first: if we take steps to increase

public trust in science, our goal should not simply be to make scientists *trusted*, we should also want them to be *trustworthy*.

In what follows, I'll first explain how recent work in the philosophy of science undermines the traditional justification given to the public for trusting science. I'll then consider two proposals that have been offered to ground public trust in science: one calling for transparency about values, the second calling for an alignment of values. I'll argue that the first proposal backfires — it rationally should *decrease* trust in science — and the second is impractical. I'll then present an alternative that is imperfect, but better than the alternatives: when scientists must appeal to values in the course of their work, they should appeal to *democratic values* — roughly, the values of the public or its representatives.

2. Trust and the Value-Free Ideal

Why should the public trust scientists? The typical answer to that question points to the nature of science. Science, it is said, is about facts, and not values. It delivers us objective, verifiable truths about the world — truths not colored by political beliefs, personal values, or wishful thinking. Of course, there are scientists who inadvertently or intentionally allow ideology to influence their results. But these are instances of *bad* science. Just as we should not allow the existence of incompetent or corrupt carpenters to undermine our trust in carpentry, we should not allow the existence of incompetent or corrupt scientists to undermine our trust in science. So long as we have institutions in place to credential good scientists and root out corrupt ones, we should trust the conclusions of science.

There is, unfortunately, one problem with this story: science isn't actually like that. In the past few decades, philosophers of science have shown that even good science requires non-epistemic value judgments. Without wading into the nuanced differences between views, I think it is fair to say that there is a consensus among philosophers of science that non-epistemic values can appropriately play a role in at

least some of the following choices: selecting scientific models, evaluating evidence, structuring quantitative measures, defining concepts, and preparing information for presentation to non-experts.[1]

These value choices can have a significant impact on the outcome of scientific studies. Consider, for example, the influential Global Burden of Disease Study (GBD). In its first major release it described itself as aiming to "decouple epidemiological assessment from advocacy" (Murray and Lopez 1996, 247). In the summary of their ten volume report, the authors describe their study as making "a number of startling individual observations" about global health, the first of which was that, "[t]he burdens of mental illnesses…have been seriously underestimated by traditional approaches… [P]sychiatric conditions are responsible…for almost 11 per cent of disease burden worldwide" (Murray and Lopez 1996, 3). Many others have cited and relied on the GBD's conclusions concerning the magnitude of mental illness globally (Prince *et al*. 2007). And nearly two decades later, the same GBD authors, in commenting on the legacy of the 1996 study, proudly noted that it "brought global, regional, and local attention to the burden of mental health" (Murray *et al*. 2012, 3).

It turns out, however, that the reported burden of mental health was driven largely by two value choices: the choice to "discount" and to "age-weight" the health losses measured by the study. Discounting is the standard economic practice of counting benefits farther in the future as being of lesser value compared to otherwise similar benefits in the present, and age-weighting involves giving health losses in the middle years of life greater weight than otherwise similar health losses among infants or the elderly. Further details about discounting and age-weighting aren't relevant to this paper; all we need to note is that the study authors acknowledged that each reflects value judgments, and that a reasonable case could be made to omit them (Murray 1996; Murray *et al*. 2012).[2] Given other methodological choices made by the authors, these two weighting functions combine to give relatively more weight to health

[1] On these points see e.g. Reiss (2017) and Elliott (2011).

[2] Indeed, in 2012 the GBD ceased age-weighting and discounting. There was also a third value choice that drove the large burden attributed to mental health: the choice to attribute all suicides to depression (Murray and Lopez 1996, 250). Because I do not know precisely how this affected the results, I set it aside here. For much more on discounting, age-weighting, and other value choices in the GBD, see Schroeder (2017).

conditions which (1) commonly affect adults or older children (rather than the elderly or young children), (2) have disability (rather than death) as their primary impact, and (3) have their negative effects relatively close to the onset or diagnosis of the condition (rather than far in the future). It should not be surprising, then, that when the GBD authors ran a sensitivity analysis to see how the decision to discount and age-weight affected the results, they discovered that the conditions most affected by these choices — unipolar major depression, anaemia, alcohol use, bipolar disorder, obsessive-compulsive disorder, chlamydia, drug use, panic disorder, post-traumatic stress disorder — were largely composed of mental health conditions (Murray and Lopez 1996, 282). Overall, the global burden of disease attributable to psychiatric conditions drops from 10.5% to 5.6%, when the results are not age-weighted or discounted (Murray and Lopez 1996, 261, 281).

I don't want to comment here on the wisdom of the GBD scientists' decision to discount and age-weight.[3] They offer clear arguments in favor of doing so and many other studies have done the same, so at minimum I think their choices were defensible. The point is that what was arguably the top-billed result of a major study — a result which was picked up on by many others, and which was still being proudly touted by the study authors years later — was not directly implied by the underlying facts. It was driven by a pair of value judgments. Had the GBD scientists had different views on the values connected to discounting and age-weighting, they would have reported very different conclusions concerning the global impact of mental illness.[4]

This case is not unique. The dramatically different assessments given by Stern and Nordhaus on the urgency of acting to address climate change can largely be traced to the way each valued the present versus the future (Weisbach and Sunstein 2009). Similar conclusions are plausible concerning the value choices involved in classifying instances of sexual misbehavior in research on sexual assault, the value

[3] I do so in Schroeder (unpublished-a).

[4] Although the sensitivity analysis was conducted by the original study authors, they do not draw any connection to their prominent claims concerning the global extent of mental illness. To my knowledge, this paper is the first to do so.

choices impacting the modeling of low-level exposures to toxins (Elliott 2011), and the value choices involved in constructing price indices (Reiss 2008).

A natural — and not implausible — response to these cases is to suggest they are outliers. Although some scientific conclusions are sensitive to value choices, the vast majority are not. The Earth really is getting warmer and sea levels really are rising, due to human activity. Vaccines really do prevent measles and really don't cause autism. These conclusions are not sensitive in any reasonable way to non-epistemic value judgments made by scientists in the course of their research. The problem, however, is that there is no clear way for a non-expert to verify this — to tell which cases are the outliers and which are not. This, I think, justifies a certain amount of skepticism. "Although some of our conclusions do depend on value judgments, trust us that *this* one doesn't," isn't nearly as confidence-inspiring as, "Our conclusions depend only on facts, not values."

I conclude, then, that rejecting the view of science as value-free, combined with high-profile examples of scientific conclusions that do crucially depend on value judgments, undermines the claim of science to public trust in a significant way. In other words, it explains why it may be rational for the public to place less trust in the conclusions of science on a broad range of issues — including in areas, such as climate change and vaccine safety, where major conclusions are not in fact sensitive to different value judgments.[5]

3. Grounding Trust in Transparency

Good science is not value-free, which undermines the standard justification given for trust in science. What, then, can scientists do to merit the public's trust? The standard response has been to appeal to transparency. If values cannot or should not be eliminated from the scientific process, scientists

[5] For similar conclusions see Douglas (2017); Wilholt (2013); Irzik and Kurtulmus (*forthcoming*); and Elliott and Resnik (2014).

should be "as transparent as possible about the ways in which interests and values may influence their work" (Elliott and Resnik 2014, 649; *cf.* Ashford 1998; Douglas 2008; McKaughan and Elliott 2018). Obviously, in order for this proposal to work, scientists would need to be aware — much more aware than most are today — of the ways in which value judgments influence their work. But, since we have independent reason to want such awareness, let us assume that calls for transparency are accompanied by a mechanism for increasing such awareness by scientists.

Would such a proposal work? Transparency about values can help ground trust in some situations, but I see no reason to think that it should broadly support public trust in science. Transparency is only useful in supporting — as opposed to eroding — trust if it enables the recipient of that information to determine how it has affected the author's conclusions. (Knowing I have a conflict of interest will typically reduce your trust in what I tell you, unless you can determine how that conflict influenced my conclusions.) Transparency, then, will only promote trust in a robust way if the public understands how value choice influenced the results, and understands what alternative value choices could have been made and how they would have influenced the results. These criteria may be satisfiable when the effect of a value choice is relatively simple. Suppose, for example, that a scientist classifies non-consensual kissing as "sexual assault", rather than "sexual misconduct", on the grounds that she believes it has more in common with rape (a clear instance of sexual assault) than it does with contributing to a sexualized workplace (a clear instance of sexual misconduct). The value judgment here is relatively simple to explain, an alternative classification is obvious, and (if the statistics involved are simple) the effect of alternative classification on the study may be relatively straightforward. So transparency could work here.

Many value choices, however, are much more complex. Think about choices embedded in complex statistical calculations — for example, those involved in aggregating climate models (Winsberg 2012) or in calculating price indices (Reiss 2008). In cases like these, it will be very hard to clearly explain the importance of any individual value choice and harder still to explain what alternative choices

could have been made. Further, many studies involve a large number of value judgments. Schroeder (2017), for example, identifies more than ten value choices which non-trivially influenced the Global Burden of Disease Study's results. Even if each of those value choices could be explained individually, it would be virtually impossible for a non-expert to figure out the interaction effects between them.

What these cases show is that even if scientists make a serious effort at transparency — not simply listing their value judgments, but attempting to explain how those judgments have influenced their results — in many cases it simply won't be possible to communicate to the public how those values have impacted their work.[6] And, if the public can't trace the impact of those values, transparency doesn't amount to much more than a warning — a reason to *distrust*, rather than to trust. A parallel realization can be seen in the way many medical schools and journals have handled researchers' conflicts of interest. Whereas in the past disclosures of conflicts of interest — essentially, transparency — were often regarded as sufficient; many have now realized that merely knowing about such conflicts does not appreciably help a reader to interpret a study. There is thus a growing move towards banning all significant conflicts of interest.[7]

4. Grounding Trust in an Alignment of Values

The previous section argued that transparency about values is not typically a solution to the problem of public trust in science. That problem, we can now see, was not caused by the fact that values were *hidden*; it was caused by the fact that the values of scientists may *diverge* from the values of any

[6] McKaughan and Elliott (2018, and in other works) suggest that scientists, through a particular sort of transparency, seek to promote "backtracking" — that is, to enable non-experts to understand how values have influenced scientists' results and to see how those results might have looked given alternative values. They seem to suggest that, at least in the cases they consider, this will frequently be possible. I am claiming that this will not generally be feasible. See Schroeder (unpublished-a) for a more detailed discussion of a particular case.

[7] See e.g. <https://ari.hms.harvard.edu/interim-policy-statement-conflicts-interest-and-commitment>

individual member of the public.[8] To promote public trust in science, then, it seems that we need to eliminate that divergence. This is the insight that motivates Irzik and Kurtulmus (*forthcoming*; *cf.* Douglas 2017; Wilholt 2013), who argue that what they call "enhanced" trust requires that a member of the public knows that a scientist has worked from value choices that are in line with her own.

If this proposal were feasible, I think it would provide a good foundation for trust. And, in certain limited cases, it may be feasible. When science is conducted by explicitly ideological organizations, members of the public may be able to make quick and generally accurate judgments about what values scientists hold, and accordingly may be able to seek out research done by scientists who share their values. (A pragmatic environmentalist, for example, might be confident that scientists employed by the Environmental Defense Fund are likely to share her values.)

Most science, however, is not conducted by explicitly ideological organizations. In these cases, it will typically be very hard for members of the public to confidently determine whether a given study relied on value judgments similar to her own. Even when this can be done (perhaps as a result of admirable transparency and clarity on the part of a scientist), it will require sustained and detailed engagement from the public, who will have to pay close attention not just to the conclusions of scientific studies, but also to their methodology. Although such close attention to the details of science would be beneficial for a great many reasons, it unfortunately is not realistic on a broad scale. There are simply too many scientific studies out there that are potentially relevant to an individual's decisions for even attentive members of the public to keep up. If our model for trust in science requires an alignment of values between the scientist and individual members of the public, trust in science can't be a broad phenomenon. Further, I don't think we want our foundation for trust in science to make that trust accessible only to those with the education and time to invest in exploring the details of individual scientific studies.

---

[8] It seems relevant to note here that distrust in science is greatest among those who identify as politically conservative, while studies show that university scientists in the U.S. overwhelming support liberal candidates for political office. Whether or not this in fact explains the distrust conservatives have in science, the argument thus far shows why such distrust could have a rational foundation.

I also — somewhat speculatively — worry that adopting this proposal would exacerbate another problem. Suppose the proposal works and, at least on some issues, members of the public are able to identify and rely upon science conducted in accordance with their own values. This, I think, might lead to a further "politicization" of science, as each side on some issue seeks scientists who share their values. Of course, once we allow a role for values in science, value-based scientific disagreement isn't necessarily a problem. Faced, for example, with one experimental design that is more prone to false positives and another that is more prone to false negatives, either choice may be scientifically legitimate. It may therefore be appropriate for more environmentally-minded citizens to rely on different studies than citizens more concerned about economic development. I worry, though, that in a culture where the public specifically seeks science done by those who share their values, it will be too easy to write off any differences in conclusions as due to value judgments — too easy for environmentalists to assume that any time pro-environment and pro-industry scientists reach different conclusions, it must be due to different underlying, legitimate value judgments. In reality, though, most such disagreements are the result of *bad* science. The worry, then, is that if we grow too comfortable with each side of an issue having its own science, it will be harder to distinguish scientific disagreements that can be traced to legitimate value judgments, from disagreements that are based on illegitimate value judgments or simple scientific error. This would be a major loss.

5. Grounding Trust in Democratic Values

I've argued that neither transparency about values nor an alignment of values can provide a broad foundation for public trust in science. Let me, then, suggest a proposal that, though imperfect, can do better. From what's been said so far, we can note a few features that a better solution should have. First, both the transparency and aligned values proposals ran into trouble because they require a great deal of attention and sophistication from the public. Most individuals simply don't have the training to

understand more technical value choices, or value choices embedded within complex calculations. And, even when such understanding is possible, it will often require a level of attention that will in practice be accessible only to the well-off. We should therefore look for a foundation for public trust which doesn't require such detailed understanding of or close attention to individual scientific studies. Second, I suggested that the aligned values proposal, in telling individuals to seek out studies conducted in accordance with their own values, could reinforce a kind of politicization that may have bad consequences. It would be better to find a proposal that wouldn't so easily divide scientists and the public along ideological lines. Third, the problem with the transparency proposal (which the aligned values proposal tried, impractically, to address) was that values, even if transparent, can be alien. In order for an individual to truly trust science, that science must be built on values that have some kind of legitimacy for her.

I think scientists can satisfy two-and-a-half of these three criteria by appealing to *democratic values* — the values of the public and its representatives — when value judgments are called for in the scientific process. The details of this proposal go beyond what I can say here.[9] But, briefly, the idea is that we look to political philosophy to tell us how to determine the (legitimate) values representative of some population. In some cases, those values might be the output of a procedure, such as a deliberative democracy exercise, a citizen science initiative, or a public referendum.[10] In other cases, it might be more appropriate to equate a population's values with the views, suitably "filtered" and "laundered", currently held by its members. ("Filtering" may be necessary to remove politically illegitimate values, e.g. racist values, and "laundering" to clean up values that are unrefined or based on false empirical beliefs.) In cases where there is a broad social consensus, that might count as the relevant democratic value; in cases where there is a bimodal distribution of values, we might say that there are two democratic values; etc.

---

[9] See Schroeder (unpublished-b) for a bit more. Many other philosophers have argued that there should be an important place for democratic values in science. See, for example, Kitcher (2011), Intemann (2015), and Douglas (2005).

[10] The extensive literature on "mini-publics" offers a promising starting point. See e.g. Escobar and Elstub (2017).

Suppose, then, that political philosophers, informed by empirical research, can give us a way of determining democratic values. I suggest that when value judgments are called for within the scientific process,[11] scientists should use democratic values when arriving at their primary or top-line results — the sort of results reported in an abstract, executive summary, or in the initial portions of the analysis. Scientists could then offer a clearly-designated alternative analysis based on another set of values, e.g. their own. I think this proposal can address two of the concerns with which I began this section, and can make some progress towards answering the third.

Let us first consider the too-much-attention and politicization problems. On the democratic values proposal, if an individual can trust that a study was competently carried out — a matter I'll return to below — then she can know, without digging into the methodological details, that its conclusions are based on objective facts plus democratic values.[12] This means that, in most cases, the public need not pay detailed attention to the methodological details of individual studies — thus solving the too-much-attention problem. Further, if scientific conclusions are based on objective facts plus democratic values, any two scientists investigating the same problem in the same social and political context should reach roughly the same conclusion. This recovers a kind of objectivity for science — not objectivity as freedom from values, but objectivity as freedom from personal biases. On this picture, the individual characteristics of a scientist should have no impact on her conclusions — a conception of objectivity that has been defended on independent grounds (Reiss and Sprenger 2014; *cf*. Daston and Galison 2007 on "mechanical objectivity"). If they are both doing good science, the environmentalist and the industrialist should reach the same top-line conclusions. And if the environmentalist and industrialist reach different

[11] This proposal is restricted to value judgments that arise within the scientific process. In particular, I do not mean for it to apply to problem selection. Scientists should be free to choose research projects that are not the projects that would be chosen by the general public. (The public, however, is under no obligation to fund such projects.) I treat the choice of research topics differently than choices that arise within the course of research because I think that scientists have different rights at stake in each case. For some related ideas, see Schroeder (2017b).

[12] There may also, of course, be methodological choices not based on non-epistemic values (including choices based on epistemic values). I set these aside here, since the problems of trust I'm concerned with don't arise in the same way from them.

top-line conclusions, it means that one or the other has made some sort of error. This, I think, provides a solution to the politicization problem: on the democratic values proposal, good science (at least in its primary analyses) will speak with a single voice.

The democratic values proposal therefore solves two of the three problems we noted above. Of course, it only does so if the public can be confident that scientists really are making use of democratic values. Why should the public assume that? Right now, I think the answer is: they shouldn't! For the democratic values proposal to work, it must be accepted by a significant portion of the scientific community, or by an easily-identifiable subset of the scientific community. If that were to happen, though, then the problem here becomes the more general one of how the public can trust scientists to enforce their own norms. The procedures and policies now in place work reasonably well, I think, to expose unethical treatment of research subjects, falsification of data, and certain other types of misconduct. I am therefore optimistic that, given a greater awareness of the role value judgments play in scientific research, a system could be devised to identify scientists who depart from a professional norm requiring the use of democratic values.

6. Science, Values, and Democracy

I've argued that the democratic values proposal can address two of the problems that faced the alternative views. But what about the third? On the transparency proposal, the values of scientists can truly be alien. If a scientist conducts research based on her own values, then, unless I happen to share those values, I have no meaningful relationship to those values. If, however, a scientists appeals to democratic values, then there is a relationship, even if I don't share those values. If democratic procedures or methods were carried out properly, then my values were an input into the process which yielded democratic values. My values are, in a sense, represented in the output of that process. This, in turn, means that those values should have a kind of legitimacy for me. In a democracy, we regularly

impose non-preferred outcomes on people when they are out-voted. So long as democratic procedures are carried out properly, this seems to be legitimate — not ideal, perhaps, but better than any available alternative. On the democratic values proposal, then, when a particular scientific conclusion is uncontested, the public can trust that that conclusion is one drawn solely from the facts, plus perhaps the values that *we* share. For most of us, who don't have the time, inclination, or ability to dig into the details of each scientific study we rely on, or who have a strong commitment to democracy, that will be enough.

I think that the foregoing provides a reasonable answer to the alien values concern. It is of course not a perfect answer. It would be better, at least from the perspective of trust, to get each member of the public access to "personalized" science conducted in accordance with her values. This, however, is impractical, as we saw when discussing the aligned values proposal. So long as that is the case, there is no way to accommodate everyone. Democratic values seem like a reasonable compromise in such a situation.

All of that said, it would be nice if we could say a bit more to those ill-served by democratic values. What should we say, for example, to an individual who knows that her values lie outside the political mainstream on some issue and is therefore distrustful of science done with democratic values on that issue? The first thing to note is that, in such cases, the democratic values proposal fares no worse (or at least not much worse) than the transparency or aligned values proposals. The democratic values proposal is fully consistent with transparency - something we have independent reason to want. So, in cases where the transparency proposal works (e.g. cases where the value choices are few, easy to understand, and computationally simple), the same advantages can be had with the democratic values proposal. Individuals who disagree with a particular value judgment and have the time and expertise to do so can determine how results would have looked under a different set of value judgments. Also, recall that I am proposing only that primary or top-line results be based on democratic values. In cases where value judgments can make a big difference — as in the Global Burden of Disease Study case discussed earlier — we might hope that scientists who hold contrary values will note the dependence of those

results on values by offering secondary, alternative analyses that begin from different value judgments. Those who have the time and expertise to dig into the methodology of scientific reports can do so, seeking out results based on values they share, as the aligned values proposal would recommend.

If the foregoing is correct, the democratic values proposal does better than the alternatives in most cases, and no worse in others. That should be sufficient reason to prefer it. But I think we can say a bit more. In what cases is the complaint from minority values most compelling? It is not, I think, when it comes from people whose values lie outside the mainstream on some issues, but within the mainstream on many other issues. The much more compelling complaint comes from people whose values consistently lie outside the mainstream — people who are consistently out-voted. Oftentimes (though of course not always) when this happens, it involves individuals who are members of groups that are or have been marginalized by mainstream society. Think, for example, of cultural or (dis)ability-based groups whose values and ways of life have been consistently treated as being less valuable and worthy of respect than the values and ways of life of the majority.

I think the democratic values proposal has two important features that can partially address such complaints. First, remember that the democratic values proposal launders and filters the actual values held by the public. Certain values — e.g. racist or sexist ones — conflict with basic democratic principles of equal worth, and so cannot be candidate democratic values. Thus, even in a racist society, telling scientists to work from democratic values will not tell them to work from racist values.[13] Second, in what I regard as its most plausible forms, democracy is not a form of government based on one person-one vote. It is a form of government based on the idea that all citizens are of equal worth and have a right to equal consideration. This suggests that, in cases where minority values are held by a group that is or has been the subject of exclusion or discrimination, democratic principles may sometimes require giving those values extra weight, or a voice disproportionate to their statistical representation in the population, as a way of accounting or compensating for past unjust treatment. Thus, democratic principles may in

[13] See Schroeder (unpublished-b) for more on this.

some cases require treating the values held by an excluded minority as democratically on a par with the conflicting values held by the majority.[14]

These considerations, I think, lessen the force of the complaint from minority values, especially in its most serious incarnation. But I don't think they eliminate it. There will still be people whose values will consistently be marginalized by the democratic view. In such cases, the main recourse available is an appeal to alternate results. If individuals with minority views can count on there being scientists who share those views, they can expect that the kind of alternative analysis they would prefer will be out there, at least in cases where it makes a difference. Of course, scientists are currently a rather homogeneous bunch along many dimensions. So this suggests that the call to work from democratic values provides (yet further) support for the importance of increasing diversity within the scientific community.[15]

[14] See Kelman (2000) for an example of this sort of argument in the context of disability.

[15] ACKNOWLEDGEMENTS TO BE ADDED

References

Ashford, Nicholas. 1988. "Science and Values in the Regulatory Process." *Statistical Science* 3.

Daston, Lorraine and Peter Galison. 2007. *Objectivity*. MIT Press.

Douglas, Heather. 2017. "Science, Values, and Citizens." In *Eppur si muove: Doing History and Philosophy of Science with Peter Machamer,* ed. Adams, Biener, Feest, and Sullivan. Dordrecht: Springer.

—. 2008. "The Role of Values in Expert Reasoning." *Public Affairs Quarterly* 22.

—. 2005. "Inserting the Public into Science." In *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*, ed. Maasen and Weingart. Dordrecht: Springer.

Elliott, Kevin. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. Oxford: Oxford University Press.

Elliott, Kevin and David Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives* 122.

Escobar, Oliver and Stephen Elstub. 2017. "Forms of Mini-Publics: an Introduction to Deliberative Innovations in Democratic Practice," NewDemocracy Research and Development Note, available at <https://www.newdemocracy.com.au/research/research-notes/399-forms-of-mini-publics>.

Funk, Cary. 2017. "Real Numbers: Mixed Messages about Public Trust in Science." *Issues in Science and Technology* 34.

Gauchat, Gordon. 2012. "Politicization of Science in the Public Sphere: A Study of Public Trust in the United States, 1974 to 2010." *American Sociological Review* 77.

Intemann, Kristin. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5.

Irzik, Gürol and Faik Kurtulmus. *Forthcoming*. "What is Epistemic Public Trust in Science?" *British Journal for Philosophy of Science*.

Kelman, Mark. 2000. "Does Disability Status Matter?" In *Americans with Disabilities: Exploring Implications of the Law for Individuals and Institutions,* eds. Francis and Silvers. Routledge

Kitcher, Philip. 2011. *Science in a Democratic Society*. Amherst, NY: Prometheus.

McKaughan, Daniel and Kevin Elliott. 2018. "Just the Facts or Expert Opinion? The Backtracking Approach to Socially Responsible Science Communication," in *Ethics and Practice in Science Communication* (eds. Priest, Goodwin, and Dahlstrom). Chicago: University of Chicago Press.

Murray, Christopher. 1996. "Rethinking DALYs." In *The Global Burden of Disease*, ed. Murray and Lopez.

Murray, Christopher and Alan Lopez (Eds). 1996. *The Global Burden of Disease*. Harvard University Press.

Murray, Christopher *et al*. 2012. Supplementary appendix to "GBD 2010: design, definitions, and metrics." *Lancet* 380.

Prince, Martin *et al*. 2007. "No health without mental health." *Lancet* 370.

Reiss, Julian. 2017. "Fact-value entanglement in positive economics." *Journal of Economic Methodology* 24.

—. 2008. *Error in Economics: The Methodology of Evidence-Based Economics*. London: Routledge.

Reiss, Julian and Jan Sprenger. 2014. "Scientific Objectivity." In *The Stanford Encyclopedia of Philosophy* (Winter 2017 edition), ed. Zalta.

Schroeder, S. Andrew. 2017. "Value Choices in Summary Measures of Population Health." *Public Health Ethics* 10.

—. 2017b. "Using Democratic Values in Science: an Objection and (Partial) Response," *Philosophy of Science* 84.

—. Unpublished-a. "Which Values Should We Build Into Economic Measures?" *Under review.*

—. Unpublished-b. "Communicating Scientific Results to Policy-makers," *manuscript on file with author.*

Weisbach, David and Cass Sunstein. 2009. "Climate Change and Discounting the Future: A Guide for the Perplexed," *Yale Law and Policy Review* 27.

Wilholt, Torsten. 2013. "Epistemic Trust in Science." *British Journal for Philosophy of Science* 64.

Winsberg, Eric. 2012. "Values and Uncertainties in the Predictions of Global Climate Models." *Kennedy Institute of Ethics Journal* 22.

PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association

Inferential power, formalisms, and scientific models

Ardourel Vincent*, Anouk Barberousse†, Cyrille Imbert§

* IHPST — CNRS, Université Paris 1 Panthéon-Sorbonne

† SND — CNRS, Sorbonne Université

§ Archives Poincaré — CNRS, Université de Lorraine

Abstract

Scientific models need to be investigated if they are to provide valuable information about the systems they represent. Surprisingly, the epistemological question of what enables this investigation has hardly been investigated. Even authors who consider the inferential role of models as central, like Hughes (1997) or Bueno and Colyvan (2011), content themselves with claiming that models contain *mathematical resources* that provide *inferential power*. We claim that these notions require further analysis and argue that mathematical formalisms contribute to this inferential role. We characterize formalisms, illustrate how they extend our mathematical resources, and highlight how distinct formalisms offer various inferential affordances.

1. Introduction. When analyzing scientific representations, philosophers of science are keen on mentioning that some models provide scientists with "mathematical resources" and "inferential power", but they seldom give a detailed analysis of these notions. This paper is devoted to the discussion of what appears to us as major mathematical resources, namely, formalisms. We thus present an analysis of the notion of formalism as well as examples from which we argue that formalisms should be acknowledged as major units of scientific activity.

We proceed as follows. In Section 2, we briefly review what philosophers of science have to say about mathematical resource and inferential power and observe that it is disappointing. In order to fill the gap we have identified, we put forward in Section 3 the three components we identify within the notion of mathematical resource. Section 4 is devoted to one of these components, namely, formalism. At last, in Section 5, we provide the reader with examples of how the choice of a formalism influences the type of knowledge scientists may draw from their representations.

2. Scientific representations and inferences therefrom. At what conditions can scientific models be used to gain information about target systems? First, a suitable semantic relation between the model and the system(s) that it stands for should obtain, so that by investigating the model, we can make legitimate inferences about its target system(s). This cannot be done unless nontrivial inferences about the model itself, as a mathematical object, can be carried out. Models are usually referred to by proper names (like "Ising model" or "Lotka-Volterra" model") or by expressions that highlight some of their mathematical properties (like "the harmonic oscillator" or "the ideal gas"). There is however more to be learnt about them than their prima facie properties. For example, solving the Ising model reveals more about Ising-like systems than their description as "sets of discrete variables representing magnetic dipole moments of atomic spins that can be in one of two states"; similarly, the mathematical content of an harmonic oscillator goes beyond "being a system that, when displaced from its equilibrium position, experiences a restoring force that is proportional to the displacement". Philosophers of science are aware of the need to investigate the epistemology of models and how we find out about concealed truths about model systems (Frigg,

2010, 257) but are surprisingly silent about how it is actually performed.[1] They are content with saying that the model is “manipulated” (Morgan and Morrison, 1997, chapter 2, passim) or that we can “play” with it (Hughes, 2010, 49), which are suggestive, but metaphoric characterizations.

Surprisingly, even accounts of applied mathematics and scientific representation that give central stage to their inferential role hardly analyze how it is fulfilled and which elements of the models contribute to it. Let us illustrate this point with Bueno’s and Colyvan’s work. They claim that “the fundamental role of applied mathematics is inferential" (Bueno and Colyvan, 2011, 352) and accordingly propose an “inferential conception" of the application of mathematics that extends Hughes’ three-step DDI account of scientific representation (see below).[2] First, a "mapping from the empirical set up to a convenient mathematical structure” (ibidem, 353) is established (immersion step); by doing so, it becomes possible “to obtain inferences that would otherwise be extraordinarily hard (if not impossible) to obtain” (ibidem, 352) (derivation step); finally, the mathematical consequences that were obtained are interpreted step in terms of the initial empirical set up (ibidem, 353) (interpretation step). Bueno and Colyvan further highlight the importance of the inferential role of mathematics for mathematical unification, novel predictions by mathematical reasoning or mathematical explanations (ibidem, 363). However, the analysis of how this inferential role is carried out shines by its absence. Bueno and Colyvan mostly analyze mathematical resources in a semantic perspective[3] and insist on the difference in content and interpretation that these make possible, e.g., when “mathematics provides additional entities to quantify

[1] Frigg, while clearly stating the problem, does not really address it and is content with briefly emphasizing the advantages of his fictional account of model concerning the epistemology of models (Frigg, 2010). As to the epistemological section of Frigg and Hartmann’s review article about scientific models, it merely points at experiments, simulations, thought-experiment as ways of investigating models (Frigg and Hartmann, 2017).

[2] Suarez’s inferential conception (Suarez, 2004) hardly addresses either the question of how inferences from models are actually carried out. For lack of space, we shall not discuss it here.

[3] Their discussion is mostly directed at the shortcomings of Pincock’s “mapping account” of the application of mathematics (Pincock, 2004).

over" (complex numbers), or is "the source of interpretations that are physically meaningful" and provide "novel prediction" about physical systems, like with the case of the interpretation of negative energy solutions to Dirac's equation (ibidem, 366).

In another paper, Bueno suggests that results are derived "by exploring the mathematical resources of the model" in which features of the empirical set up are immersed (Bueno, 2014, 379, see also 387) and that results emerge "as a feature of the mathematics" (ibidem) or by using "the particular mathematical framework" (ibidem, 385). What this inferential power of mathematics should be specifically ascribed to remains unclear. Bueno and Colyvan (2011, 352) just claim that the "embedding *into a mathematical structure* makes it is possible to obtain inferences". They also emphasize how, with the help of appropriate idealizations, "the *mathematical model* [can] directly [yield] the results" (ibidem, 360, our emphasis). But elsewhere in the paper, consequences are said to be drawn "*from the mathematical formalism*, using the mathematical structure obtained in the immersion step" (ibidem, 353, our emphasis).

What are we to make of these various claims? A prima facie plausible answer to this question might be that structures and formalisms are the two sides of a same inferential coin. However, this answer is not satisfactory, since, as is well-known, mathematical structures can be presented in different formalisms, which, as we shall see in Section 4, are associated with different inferential possibilities. Another blind spot in Bueno's and Colyvan's account is that while the derivation step is claimed to be "the *key point* of the application process, where consequences from the mathematical *formalism* are generated" (ibidem, 353), the question of how inferences are drawn with the help of formalisms is left under-discussed.

We draw from this brief analysis of Bueno's and Colyvan's views that the notions of mathematical resource and inferential power, which are commonly used when discussing applications of mathematics, are often mere labels in need of further investigation. Coming back to the seminal ideas presented by Hughes and extended by Bueno and Colyvan is of little help because Hughes' paper lacks precise answers to the following precise questions: What are exactly mathematical resources? What is their inferential power? In his DDI (Denotation, Demonstration, and Interpretation) account of scientific representation, Hughes claims that scientific representations have an "internal dynamic", whose effects we can examine (1997, 332), and "contain *resources* which enable us to demonstrate the results we are interested in". A general notion of resource is appropriate to capture the variety of ways in which demonstrations can be

carried out; however, the claim that the deductive power comes from "the *deductive resources* of mathematics they employ" (ibidem, 332) is too vague and is left unanalyzed.

3. Components of mathematical resources. How are the notions of inferential power and mathematical resources to be analyzed? Are they linked to structures or to symbolic systems and formalisms? In this section, we claim that formalisms are an important component of the notions of inferential power and mathematical resource and should be analyzed in their own right.

Let us begin by briefly presenting what are, according to us, the three main components of the notions of mathematical resource and associated inferential power. First, mathematical structures, *to the extent that they are tractable*, are undoubtedly an important part of the mathematical resources that are used in mathematical modeling. As argued by Cartwright, theories are no "vending machines" that "drop out the sought-for representation" (1999, 247); scientific models are no vending machines either and scientists must make the best of the models that they know to be tractable. Accordingly, the content of models often needs to be adapted by means of idealizations, approximations (Redhead 1980), abstractions, by squeezing representations into the straightjacket of a few elementary models (Cartwright, 1981), or by drawing, from the start, on the pool of existing tractable models (Humphreys, 2004, Barberousse and Imbert, 2014).

Second, mathematical knowledge associated with structures is also to be counted as a distinct mathematical resource, which allows for new inferences when it is available. Let us take the well-known example of Koenigsberg's seven bridges. The impossibility of crossing them once and only once in a single trip can be demonstrated by applying a result from graph theory. Similarly, the explanation of the life-cycle of the Magicicada (Baker 2009, Colyvan 2018) is provided by the application of a number-theoretic property of prime numbers to life-cycles of species.

At last, formal settings or formalisms provide languages in which theories are developed, calculations carried out, and inferences drawn from models. Examples of formalisms are Hamiltonian formalism, path integrals, Fourier representation, cellular automata, etc. We provide a detailed analysis of some of these below. Contrary to mathematical structures, formalisms are partly content neutral (though form and content are often intertwined in scientific representations). As providing a partially stan-

dardized way of making inferences, they are important tools for scientists, which in turn justifies considering them as important units of analysis in the philosophy of science. Other authors have started exploring the idea that format matters in scientific activities. Humphreys gives general arguments to this effect and emphasizes the difference between formats that are appropriate for human-made and format that suit computational inferences (2004). Vorms (2009) also emphasizes the general importance of formats of representation when toying with theories or models. Formalisms are a specifically mathematical type of format whose role needs further investigation. This is what we do in the next section.

4. What are formalisms? As briefly stated above, formalisms are mathematical languages that allow one to present mathematical statements or objects and draw inferences about them by means of general inference rules. For example, *Hamiltonian formalism* is one of the formalisms through which scientists may find out means to solve differential equations. *Path integrals* is another formalism of this kind, with the help of which one may also solve (partial) differential equations. Let us illustrate the latter point further: the integral solution of the Schrödinger equation requires using a mathematical object, the *propagator,* whose calculation the path integrals formalism makes easier. *Fourier representation or formalism* enables one to represent mathematical functions as the continuous sum of sine functions (or complex exponential functions), so that harmonic analysis, i.e. the decomposition of a signal in its harmonic frequencies, may be performed. It also provides modelers with a way to express the solutions of some partial differential equations, such as the heat equation. Finally, formalisms like *numerical integrators, cellular automata, lattice Boltzmann methods,* and *discrete variational integrators*, are indispensable in current computational science.

Formalisms consist in the following elements:

i. elementary symbols;

ii. syntax rules that determine the set of well-formed expressions;

iii. inference rules;

iv. a partly detachable interpretation, both mathematical and physical.

Their use is facilitated by

v. translation rules that indicate how to shift from one formalism to another.

Let us illustrate these elements by discussing in more detail the above examples. In the Hamiltonian formalism, elementary symbols are used for a variable and its conju-

gate momentum: “(q, p)”, or for Poisson brackets ‘’{.,.}’'. Among the syntax rules that are specific to Hamiltonian formalism, some allow one to rewrite Hamilton equations by using the canonical variables. Inferences rules allow the users to use action-angle variables (I, *theta*) and to solve equations by using these coordinates because this change of variables opens the possibility to deal with integrable systems, thus providing a systematic method to solve *exactly*, i.e., in closed forms, differential systems like the simple pendulum, and more generally, any 1D-conservative system. Indeed, due to this change of variables, one takes full advantage of the existence of conserved quantities in mechanical systems, which are then used as variables (actions) in Hamilton equations. This allows constructing the solution of the equations by “quadrature’' (Babelon et al. 2003, chapter 2). An example of a translation rule is the Legendre transform that allows one to shift to Lagrangian formalism. Similarly, in the case of Fourier transforms, an elementary specific symbol is $f\hat{}$, which corresponds to the Fourier transform of the function *f*. Scientists use sets of rules that describe the Fourier transforms of some typical functions, such as the constant function, the unit step function, and the sinusoids, but also rules for the convolution product, viz. the Fourier transform of the convolution f o g is the product of Fourier transforms of f and g: $(f \text{ o } g)\hat{} = f\hat{}.\ g\hat{}$, so that solutions of equations may be found within Fourier space. An inverse Fourier transform is also defined, which enables one to move back from the Fourier transform $f\hat{}$ to the function f (this is again a translation rule).

As emphasized above, formalisms are (partly) content neutral and thus “exportable”, even though they usually come with a privileged physical interpretation. As a matter of fact, most formalisms have been developed within a peculiar modeling context or are linked to a physical theory. From this origin, the most successful ones may become autonomous and depart from their original, physical interpretation. For example, Hamiltonian formalism was initially developed in the context of classical mechanics but is nowadays autonomous and used in other physical contexts. Path integrals originally come from the study of Brownian motion (Wiener 1923) and quantum mechanics (Feymann 1942) but are currently used in other fields like field theory and financial modeling.

The mathematical interpretation of formalisms may sometimes be detachable. For example, the transition rules associated with cellular automata (see below) do not have any obvious mathematical interpretation. Further, although some formalisms are linked to acknowledged mathematical theories (e.g., the Fourier formalism is linked to

the theory of complex functions), they differ from genuine mathematical theories, as shown by the example of path integrals, in which the formalism is used in the absence of any uncontroversial mathematical theory that could back it up. The definition of a path integral:

$$K(b,a) = \int_a^b e^{\frac{2i\pi}{h}\int_{t_b}^{t_a} L\,dt}\, Dx(t)$$

requires using a measure "Dx'', to which no general, rigorous definition can be given yet. This mathematical concern does not prevent physicists from using path integrals anyway, as testified by the following quote: "The question of how the path integral is to be understood in full generality remains open. Given this, one might expect to see the physicists expending great energy trying to clarify the precise mathematical meaning of the path integral. Curiously, we again find that this is not the case" (Davey 2003, 450).

Let us finally emphasize that formalisms also differ from formulations of physical theories and allow philosophers of science to address different philosophical problems. Formulations of theories, in particular axiomatic ones, are explored when questions about conceptual content and metaphysical implications are raised. They pertain to foundational issues. Whether a given formulation involves calculus is a peripheral issue in this context. By contrast, the primary virtue of a formalism is to allow modelers to draw actual inferences from a theory or model. The inferential rules it contains are more important than the mathematical rigor of the language in which it is expressed.

5. Choosing a formalism. So far, we have argued that the inferential power that is required to explore models is partly brought about by formalisms, and we have given examples thereof. Accordingly, formalisms have to be carefully examined by philosophers of science if they are to provide a fine-grained analysis of how scientific knowledge is produced in practice. We now aim to show that there is no unique description of formalism-rooted inferential power since different formalisms allow for different types of inferences and are adapted to different types of inquiries. We do so by providing examples of these differences and of the factors that guide scientists when choosing the formalism that is best suited to the task at hand.

How do scientists decide which formalism to use in a given inquiry? The choice may first depend on the type of models at hand. For example, the path integral formalism is

well adapted to solve systems with many degrees of freedom (Zinn-Justin 2009) and makes "certain numerical calculations in quantum mechanics more tractable" (Davey 2003, 449). Lagrangian formalism offers a well-suited framework to solve equations describing constrained systems (Goldstein 2002, 13, Vorms 2009, 15). Fourier representation allows one to solve, e.g., the differential equations describing the time evolution of electrical quantities in networks. In this case, differential equations are transformed into *algebraic equations* on variables in Fourier space, which may be easier to solve. Finally, with the change of action-angle variables, Hamiltonian formalism potentially provides exact solutions for integrable systems, which have as many independent conserved quantities as degrees of freedom.

The use of a particular formalism is also guided by epistemic goals. Depending on the chosen formalism, different kinds of properties, general (e.g. periodicity, symmetry) or particular (dynamical), may be inferred from the same model. Let us illustrate this point with the example of prey-predator models in ecology. Among these, some obey Lotka-Volterra (LV) equations and represent transforming populations with a system of two coupled equations. If they are investigated within the Hamilton formalism, *general properties* of these models can be found without setting initial conditions or numerical values for the involved parameters. The reframed models can indeed be shown to be integrable, like the simple pendulum in classical mechanics. Dutt explicitly emphasizes the advantages of using this formalism for a two-species LV system:

> "In dealing with the problems involving *periodicity*, the Hamilton-Jacobi canonical theory has a distinct advantage over the conventional methods of classical mechanics. In this approach, one introduces action and angle variables through canonical transformations in such a way that the angle variable becomes cyclic. One then obtains the frequency of oscillation by taking the derivative of the Hamiltonian with respect to the action variable. One may thus *bypass the difficulty* in obtaining the complete solutions of the equations of motion, *if these are not required*." (Dutt, 1976, 460, our emphasis)

LV models can also be solved with the help of computers and generic numerical integrators when the aim is to obtain particular dynamics for specific values of parameters and initial conditions. Such numerical solutions of the LV model can also be provided by specific formalisms, such as discrete variational integrators (Krauss 2017, 34; Tyranowski 2014, 149). In that case, discrete equations are derived from a discrete least action principle, which is well-suited to conservative systems, like the LV sys-

tem. Discrete variational integrators allow for the preservation of general properties like the conservation of global quantities, viz. energy, momenta, and symplecticity. This discrete formalism comes with mathematical constraints on the discretization of time since the time step has to be adaptive in order to guarantee the conservation of global quantities (Marsden & West 2001, Section 4.1).

Finally, let us mention that LV models can also be studied by using *cellular automata* (CA) and associated formalism, with the following advantages:

> [a rather general predator-prey model] is formulated in terms of automata networks, which describe more correctly the *local character* of predation than differential equations. An automata network is a graph with a discrete variable at each vertex which evolves in discrete time steps according to a definite rule involving the values of neighboring vertex variables. (Ermentrout and Edemstein-Keshet 1993, 106)

On the one hand, CA are discrete dynamical systems, but on the other, they are also a nice means to practice science with the help of a computationally simple formalism (in terms of transition rules). They can be extremely powerful. For example, rule 110 is Turing complete and, like lambda-calculus, can emulate any Turing machine and therefore complete any computation. In contrast with the case of Hamilton formalism, CA-based inferences from prey-predator models are carried out for specific values and parameters. As CA are described by local rules, these inferences merely pertain to local variations in the model. However, the simplicity of these rules is a tremendous advantage for modeling and code-writing. For instance, CA allow one to easily add rules for the pursuit and evasion of populations as well as rules for age variation (Boccara et al. 1993, Ermentrout and Edemstein-Keshet 1993, see also Barberousse and Imbert 2013 for an analysis of CA as used in fluid dynamics and compared with Navier-Stokes based methods).

Let us now turn to a different example illustrating how different the epistemological effects of using this or that formalism may be. Crystals are currently modeled as lattices that come under two forms, *lattices in real space* and *lattices in reciprocal space*. Each is associated with a specific formalism. Within the *real space lattice* formalism, crystals are described with a vector R expanded on a vector basis (a_1, a_2, a_3) which corresponds to crystal directions, and *alpha*, *beta*, *gamma* are the corresponding angles. Inferences about *symmetry* of crystals are usually made within this type of representation since the real space is well adapted to studying discrete translations and rotations.

Crystals can also be described with the help of a vector R* in a *lattice in reciprocal space*. There is a clear correspondence between the two spaces since they are dual. Given R in the real space, we can derive R* in the reciprocal space, and conversely. The two spaces are related by a Fourier transform. However, the *reciprocal space* can be more convenient because inferences about *diffraction and interference patterns* are easier to carry out in the Fourier representation. As stressed by Hammond in a textbook of crystallography:

> the reciprocal lattice is the basis upon which the geometry of X-ray and electron diffraction patterns can be most easily *understood* and […] the electron diffraction patterns observed in the electron microscope, or the X-ray diffraction patterns recorded with a precession camera, are simply sections through the reciprocal lattice of a crystal (Hammond 2009, 165).

This example shows that facilitating inferences may have various epistemological effects. Some are relevant to computational aspects and the predictions or explanations that scientists are able to produce in practice. Others pertain to the way scientists understand and reason about models and their target systems. This example also shows how different epistemic goals (symmetry-oriented vs. interference-oriented investigations of crystals) determine which formalism is chosen.

Overall, the above shows that formalisms not only have an important impact on the amount of results scientists may produce, but also on the types of results that are attainable. The examples we have discussed also highlight that the existence of a variety of formalisms is a source of epistemic richness and enhanced inferential power for scientists because it provides them with multiple ways of investigating the same mathematical structures or structures that are related by suitable morphisms.

6. Conclusion. The above proposals are meant to contribute to the epistemological question of what provides models with inferential power and helps scientists succeeding in their inquiries. We have shown that some of this inferential power is brought about by the formal symbolic tools that scientists use to present and investigate mathematical models. Our second claim is that all formal settings do not enable the same types of inferences nor are suited to all epistemic goals. Accordingly, a fine-grained analysis of the conditions of scientific progress needs, among other things, to focus on formalisms.

Our epistemological analysis is not tied to any particular theory of scientific representation. However, by showing that inferences actually hinge on choice of formalism, it suggests that a theory of scientific representation that is cashed out in terms of structures is too abstract to account for the various ways equations are solved in practice and information extracted from scientific models.

References

Babelon Olivier, Bernard Denis, and Talon Michel. 2003. *Introduction to classical integrable systems*, Cambridge: Cambridge University Press.

Baker, A. 2009. "Mathematical Explanation in Science". *British Journal for the Philosophy of Science* 60 (3): 611–633.

Barberousse, Anouk, and Cyrille Imbert. "New Mathematics for Old Physics: The Case of Lattice Fluids." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 44 (3) : 231–41.

Barberousse, Anouk, and Cyrille Imbert. 2014. "Recurring Models and Sensitivity to Computational Constraints" *The Monist* 97 (3): 259–79.

Boccara Nino, Roblin O. and Roger Morgan. 1994. Automata network predator-prey model with pursuit and evasion, *Physical Review E* 50 (6): 4531–41

Bueno, Otávio, and Mark Colyvan. 2011. "An Inferential Conception of the Application of Mathematics". *Noûs* 45 (2): 345–74.

Bueno, Otávio. 2014. "Computer Simulations: An Inferential Conception". *The Monist* 97 (3): 378–98.

Cartwright, Nancy (1983). *How the Laws of Physics Lie*. Clarendon Press, Oxford.

Cartwright, Nancy.1999. "Models and the Limits of Theory: Quantum Hamiltonians and the BCS Models of Superconductivity". In *Models as Mediators*, ed. Mary S. Morgan and Margaret Morrison Morgan, Cambridge: CU Press: 241–81.

Colyvan, Mark. Forthcoming. "The Ins and Outs of Mathematical Explanation", *Mathematical Intelligencer*.

Davey Kevin. 2003. “Is Mathematical Rigor Necessary in Physics?” *The British Society for the Philosophy of Science*, 54(3): 439–463

Dutt Ranabir. 1976. “Application of the Hamiltonian-Jacobi Theory to Lotka-Volterra Oscillator”, *Bulletin of Mathematical Biology*, 38: 459–465.

Ermentrout G. Bard and Edemstein-Keshet, Leah. 1993. “Cellular Automata Approaches to Biological Modeling”. *Journal of Theoretical Biology* 160: 97–133.

Feynman, Richard. P. 1942. “The Principle of least action in quantum mechanics”, *PhD. diss.*, Princeton University.

Frigg, Roman. 2010. “Models and Fiction”. *Synthese* 172 (2): 251–68.

Frigg, Roman, and Stephan Hartmann. 2017. “Models in Science.” In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/models-science/.

Goldstein, Herbert. 2002. *Classical Mechanics*. Reading, Mass: Addison-Wesley.

Hammond, Christopher. 2009. *The Basics of Crystallography and Diffraction*, Oxford University Press.

Hughes, Robert I.G. 1997. “Models and Representation”. *Philosophy of Science* (Proceedings): 64: S325–S336.

Hughes, Robert I.G. 2010. *The Theoretical Practices of Physics: Philosophical Essays*. Oxford: Oxford University Press.

Humphreys, Paul. 2004. *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Kraus, Michael. 2017. “Projected Variational Integrators for Degenerate Lagrangian Systems”, preprint: https://arxiv.org/pdf/1708.07356.pdf

Marsden Jerrold E. and West Matthew. 2001. “Discrete Mechanics and Variational Integrators”, *Acta Numerica*, 10: 357–514.

Morgan, M., and Margaret Morrison (1999). *Models as Mediators*. Cambridge University Press.

Pincock, Christopher. 2004. "A new perspective on the problem of applying mathematics", *Philosophia Mathematica* 3 (12), 135-61.

Redhead, M. 1980. "Models in Physics", *The British Journal for the Philosophy of Science*, 31(2): 145-163

Suarez, Mauricio. 2002. "An Inferential Conception of Scientific Representation", *Philosophy of Science* 71 (5): 767-779

Tyranowski Tomasz. M. 2014. "Geometric integration applied to moving mesh methods and degenerate Lagrangians". Ph.D. diss., California Institute of Technology.

Vorms, Marion. 2011. "Formats of Representation in Scientific Theorizing." In *Models, Simulations, and Representations*, edited Paul Humphreys and Cyrille Imbert. Routledge.

Wiener, Norbert. 1923. "Differential space". *Journal of Mathematical Physics* 2: 131-174.

Zinn-Justin Jean. (2009), Path Integral, *Scholarpedia*, 4(2): 8674.

Representation Re-construed: Answering the Job Description Challenge with a Construal-based Notion of Natural Representation

Abstract: Many philosophers worry that cognitive scientists apply the concept REPRESENTATION too liberally. For example, William Ramsey argues that scientists often ascribe natural representations according to the "receptor notion," a causal account with absurd consequences. I rehabilitate the receptor notion by augmenting it with a background condition: that natural representations are ascribed only to systems construed as organisms. This Organism-Receptor account rationalizes our existing conceptual practice, including the fact that scientists in fact reject Ramsey's absurd consequences. The Organism-Receptor account raises some worrying questions, but as a more faithful characterization of scientific practice it is a better guide to conceptual reform.

Abstract: 100 words

Total: 4,995 words

1. Introduction. There is a common complaint among philosophers that scientists use the word "representation" too liberally. Representation is often contrasted with indication: representation is a distinction achieved by maps, linguistic performances, and thoughts, whereas indication is a less-demanding state achieved by thermostats, which indicate ambient temperature, and refrigerator lights, which indicate whether the door is open (Dretske 1981; Cummins and Poirier 2004). However, cognitive scientists often ascribe representations when it seems that mere indication is all that is called for. We commonly say that hidden layers in a neural network represent concepts, or that neurons in V1 represent visual edges, because they reliably respond differently to the circumstances they are said to represent (Ramsey 2007, 119–20; cf. Hubel and Wiesel 1962). But these "representations" are thin-blooded compared to paradigmatic conventional representations. For example, they cannot be invoked in the absence of an appropriate stimulus. So are cognitive scientists conceptually confused? Do they exaggerate their claims? And if the natural representations posited by cognitive scientists aren't genuine representations, is the cognitive revolution dead?

William Ramsey provides an excellent book-length exploration of these worries, articulating a qualified pessimism about their answers:

> …we have accounts that are characterized as "representational," but where the structures and states called representations are actually doing something else. This has led to some important misconceptions about the status of representationalism, the nature of cognitive science and the direction in which it is headed. (2007, 3)

Ramsey describes the "job description challenge": to give an account of the distinctive properties of representations in virtue of which appealing to them serves a special

explanatory role. If the job description challenge can be met, then we can formulate a plan for conceptual reform.

I undertake Ramsey's challenge, but with a metadiscursive twist: I describe the Organism-Receptor account, which articulates conditions for ascribing representations, in virtue of which such ascriptions achieve a special explanatory purpose. The account is merely suggestive about the properties that distinguish first-order representational states from non-representational states; it says more about the mental state of the ascriber than about the representation-bearing system. However, the Organism-Receptor account provides a more adequate characterization of scientists' practice than Ramsey's.

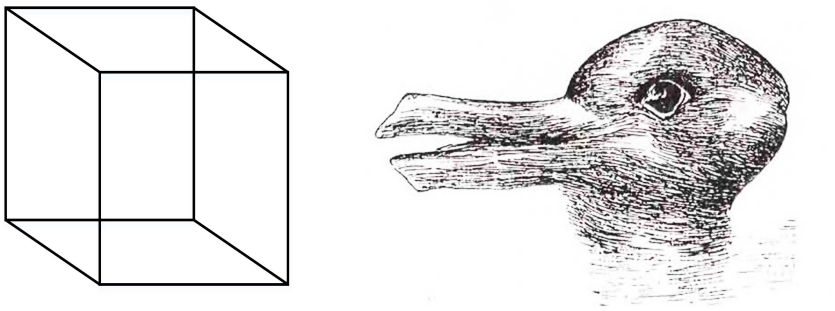
My main aim in this paper is to push back against pessimistic evaluations of the existing practice of representation-ascription in cognitive science, like Ramsey's. I will focus on Ramsey's critique of the "receptor notion," a flawed causal theory of representation that he attributes to some cognitive scientists. Ramsey argues that the receptor notion has absurd consequences, although scientists do not accept them. By augmenting the receptor notion with a construal-based background condition, I can explain why scientists do not draw these absurd conclusions. Whereas Ramsey's pessimistic account of scientists' practice of ascribing representations finds it wanting and is extensionally inadequate, mine rationalizes our extant conceptual practice (though that practice is not beyond criticism). I conclude that my apologetic account is a more charitable and adequate interpretation of existing scientific practice than Ramsey's.

2. Ramsey on the "Receptor Notion." Ramsey argues that natural representations in cognitive science are often ascribed according to the "receptor notion," a crude causal theory of representation. According to the receptor notion, a state *s* represents a state of affairs *p* if *s* is regularly and reliably caused by *p* (2007, 119).

Ramsey claims that the receptor notion is what justifies the ascription of representations to cells in V1 that detect visual edges, cells in frog cortex that detect flies, and the mechanisms in Venus flytraps that cause their "jaws" to close (119–23). Ramsey argues that this receptor notion is too liberal to be useful to scientists. For example, it is susceptible to the "disjunction problem" (Fodor 1987): since frog neurons respond reliably to visual stimulation by flies *or* (say) BBs, we should say that the content of the representation is *fly-or-BB*, rather than *fly*. Likewise, Venus flytraps represent objects in a particular range of sizes rather than *edible insects*, and the human concept GOAT represents *goats-or-weird-looking-sheep*. Such disjunctive content-ascriptions are usually considered absurd. Absent a clever fix, we must embrace unwieldy, disjunctive contents for representations or we must reject the receptor notion (Ramsey, 129).

Dretske's (1988) teleofunctional theory of representation is a sophisticated twist on the receptor notion that avoids the disjunction problem. On Dretske's view, a representational state must not only be causally dependent on the state of affairs it represents, but must serve a function for its containing system in virtue of this causal dependency. This extra condition motivates constraints on representational content that eliminate problematic disjunctive contents. Dretske's theory is subject to some subtle criticisms that I will discuss in Section 6, but the Organism-Receptor account will preserve some of the teleological character of Dretske's theory.

Ramsey's most compelling objection to the receptor account, including Dretske's sophisticated version, is that it justifies ascribing representational contents to states that are not, in fact, representational: smoke "represents" fire since the latter causes the former. Likewise, the firing pin of a gun "represents" whether the trigger is depressed, and rusting iron "represents" the presence of water and oxygen (138–47). Ramsey claims, plausibly, that these are absurd consequences. I find Ramsey's reductio



Ambiguous figures. Left: The Necker cube. Right: The duck-rabbit (image from Jastrow 1899).

compelling, but reject a different premise than he does. Rather than conclude that cognitive scientists have a bad conceptual practice, I question whether his characterization of the receptor notion is a charitable understanding of what happens in cognitive science. After all, cognitive scientists do not generally claim that GOAT denotes *goats-or-sheep* (at least for competent judges of goathood), or that firing pins represent anything.

3. A Construal-based Notion of an Organism. I argue that something like the receptor notion can be salvaged if being a receptor is contextualized in terms of construal. Construal (also called "seeing-as") is a judgment-like attitude whose semantic value can vary licitly independently of the state of affairs it describes. For example, we can construe an ambiguous figure like the Necker cube as if it were viewed from above or below, or the duck-rabbit as if it were an image of a duck or of a rabbit (Roberts 1988; see also Wittgenstein 1953). We can construe an action like

skydiving as brave or foolhardy, depending on which features of skydiving we attend to.

On a construal-based account of conceptual norms, a concept (e.g. REPRESENTATION) is ascribed relative to a construal of a situation. For example, perhaps I fear something only if I construe it as dangerous to me or detrimental to my ends (Roberts 1988). Daniel Dennett's (1987) intentional stance is a more familiar example: according Dennett, a system has mental states if and only if we construe it in such a way that its behavior is explainable in terms of a belief-desire schema.

I propose that construing something as an organism involves construing it such that it has goals and behavior, and believing that it has mechanisms that promote those goals by producing that behavior. More precisely:

> **Organism-Construal.** A subject *a* construes a system *x* as an organism in a context[1] *c* if and only if, in *c*,
>
> (O1) *a* attributes a set of goals *G* to *x*,
>
> (O2) *a* attributes a set of behaviors *B* to x,
>
> (O3) *a* believes that the elements of *B* function to promote elements of *G*,
>
> (O4) *a* believes that *x* possesses a set of mechanisms *M*, and
>
> (O5) *a* believes that the elements of *M* collectively produce the elements of *B*.

My main argument does not rely on all the details of Organism-Construal; it could be replaced by a different explication of what it is to see something as an organism. But Organism-Construal captures an intuitive notion of a critter. First of all, we normally take living critters to have goals, such as survival and reproduction, and behaviors that

[1] The relevant notion of a context is something like MacFarlane's (2014) "context of assessment."

promote those goals. However, Organism-Construal does not require that an organism really have goals (whatever that involves) or exhibit behavior (however that's distinguished from other performances). To see something as an organism according to Organism-Construal, the construing subject need only *attribute* goals to the system, and see some of its performances as behaviors that promote those goals. Such goals could include relatively specific aims such as locating food, getting out of the rain, or driving home. We sometimes also attribute goals and behaviors to non-living things, such as automated machines. For example, we might say that a robot vacuum has the goal of cleaning the floor, which it accomplishes by sucking up dust. Or I might say that my GPS navigation computer is trying to kill me, which it accomplishes by consistently giving me directions that lead me through strange, dangerous backroads. Condition (O3) is expressed in terms of belief instead of attribution, meaning that the construing subject must sincerely believe that an organism's putative behaviors function to promote its putative goals. When and insofar as someone construes a system in this way, the conditions (O1)–(O3) above are satisfied.

Conditions (O4)–(O5) require that the system's behavior be explainable by appeal to mechanisms. "Mechanisms" here should be understood in roughly the sense meant by the new mechanists (Machamer, Darden, and Craver 2000; Bechtel and Abrahamsen 2005; Craver 2007): organized structures of component parts and operations that produce a phenomenon, and the description of which is an explanatory aim of some scientific projects. Much explanation in biology and neuroscience plausibly follows a mechanistic model, and likewise in cognitive science. Daniel Weiskopf (2011) has argued that cognitive explanations are not properly mechanistic, but even on his view cognitive explanations are extremely similar to mechanistic ones, distinguishable only because the relationship between components of cognitive models and their physiological realizers is relatively opaque. Regardless, cognitive scientists use the word "mechanism" to refer to the referents of their models,

just as biologists and neuroscientists do. I am more moved by the similarities between the biological and the cognitive sciences than the differences. Therefore, like Catherine Stinson (2016), I acknowledge Weiskopf's concerns but nevertheless adopt the language of "mechanisms."

Not all of a system's mechanisms function to produce behavior. For example, biological organisms have metabolic and other mechanisms that maintain bodily integrity. Such mechanisms may need to function correctly as a background condition for the organism to behave, but scientists do not typically take behavioral patterns to be the explanandum phenomena of such mechanisms. Let us call mechanisms that do contribute to the explanation of behavior *behavioral mechanisms*. As for what it means for a system to "possess" a mechanism, a mereological criterion will do for now: the mechanism must be a part of the system. Condition (O5) is meant to limit the mechanisms in the set *M* to behavioral mechanisms.

So far so abstract; let's consider an example. The robot Herbert was designed to wander autonomously through the MIT robotics lab, avoiding obstacles, and collecting soda cans with its arm (Brooks, Connell, and Ning 1988). Herbert can be construed as an organism, even though it is not alive, as long as one (O1) attributes goals, like avoiding collisions and collecting soda cans, to Herbert, (O2) sees some of Herbert's performances as behaviors, (O3) believes that Herbert's behaviors promote its goals, and (O4) believes that Herbert possesses mechanisms that (O5) explain its behavior. Herbert does possess mechanisms for accomplishing goals; it is equipped with sensors, computers, and motors that coordinate its locomotion and its grasping arm. And most people readily anthropomorphize Herbert enough to see it as a goal-directed, behaving system (pace Adams and Garrison [2013], who insist that Herbert has its designers' goals, but no goals of its own). Anyone willing to engage in the imaginative attribution of goals and behavior to Herbert can see Herbert as an organism, even if on reflection they believe Herbert is not literally an organism. The

willingness to ascribe representations to a system plausibly waxes and wanes along with one's willingness to construe the system as an organism in something like the sense described above. There are psychological limits on the willingness to attribute goals and behaviors to systems relatively unlike animals, and these limits may vary between individuals.

4. The Receptor Notion Re-construed. Returning now to the receptor notion of natural representation, I suggest that it can be augmented in the following way:

> **Organism-Receptor.** A state *s* represents a state of affairs *p* if
>
> (R1) *s* is regularly and reliably caused by *p*, and
>
> (R2) *s* is a functional state of a behavioral mechanism possessed by an organism.

Organism-Receptor is not a construal-based explication, but it depends on a construal-based account of ORGANISM. It preserves the spirit of Ramsey's receptor notion, with the added condition that representations be ascribed to parts of systems construed as organisms. Representation-ascriptions guided by Organism-Receptor inherit their plausibility from the plausibility of the corresponding construal of some system as an organism. Most accounts of cognitive representation require there to be a representational subject of some kind (e.g. Adams and Aizawa 2001; Rupert 2009; Rowlands 2010), and on Organism-Receptor the organism serves this role. We can constrain the acceptable contents of these representations by requiring they correspond to descriptions of *p* according to which *p* is relevant to the pursuit of an organism's goals. This appeal to goals is not ad hoc, since according to Organism-Receptor representations are ascribed to organisms, i.e. systems to which we've already attributed a set of goals. Thus, like Dretske's (1988) and Millikan's (1984)

teleofunctional accounts, this construal-based account addresses the disjunction problem by appealing to goals of organisms.

The metadiscursive job-description challenge is to provide criteria of ascription for representations, in virtue of which representation-ascriptions achieve some explanatory purpose. I have provided criteria of ascription, so what is their purpose? On Donald Davidson's (1963, 5) account of intentional action, actions are performed under the guise of a privileged description (or set of descriptions). Davidson flips the light switch in order to turn on the light, but not in order to alert the prowler outside (whose presence is unknown to Davidson) that he is home, though he also does the latter. Davidson calls this feature of action its "quasi-intensional character." Behavioral mechanisms also have something like a quasi-intensional character, since there are privileged descriptions that make explicit how they and their components contribute to an organism's capacity to pursue its goals. For example, edge-detecting cells in V1 fire in order to identify boundaries in an organism's environment, not to consume glucose, though they also do the latter. The use of representation-talk by cognitive scientists, as licensed by Organism-Receptor, is a way to habitually mark these privileged descriptions and distinguish them from other descriptions of the same states or events. And since cognitive science is concerned with the functional structure of behavior-coordinating mechanisms rather than other features of cognitive systems, it is easy to see why representation—even in this relatively thin sense—has always been the dominant theoretical perspective in cognitive science. This focus on quasi-intensional characterization may even be what makes the cognitive scientific perspective distinctive (on scientific perspectives, see e.g. Giere 2006).

The Organism-Receptor account provides us with resources to salvage the receptor notion from Ramsey's reductio. It is plausible to suppose that cognitive scientists generally ascribe natural representations to systems against an imaginative

background like this. After all, most cognitive science concerns the mechanisms of living systems, especially animals (except in computer science and some computational modeling, where the object of attention is a formal object like a connectionist network that is presumed to be analogous in some way to such a mechanism). Such systems are easily construed as organisms in the sense of Organism-Construal. Non-living things and even non-animals are in general more difficult to construe as organisms in that sense, since they are often perceived to lack goals, the capacity to behave, or both.

5. The Organism-Receptor Notion in Context. Consider a strong case of representation, like fly-detecting cells in frog visual cortex. We construe frogs as systems that exhibit goal-directed behavior and believe they possess mechanisms that explain that behavior. Frog visual cortex contains mechanisms that (along with other mechanisms) explain behaviors like fly-catching. When we identify cells in frog visual cortex that fire in response to the visual presence of flies (or fly-like objects), we ascribe representational properties to those cells. The contents we ascribe to representations in frog visual cortex are constrained by the goals we attribute to frogs. *That a small insect is present* is a suitable content because flies can be consumed for energy; *that a wiggly BB is present* does not have this significance for frogs, although BBs may be indistinguishable from insects by the mechanisms in the frog's visual cortex. Nevertheless, the relationship between fly-presence and the frog's goals provide a ground for privileging non-disjunctive descriptions of representational content.

The Organism-Receptor account also explains why liminal cases of representation, like the case of Herbert, are liminal. We can say that Herbert represents such states of affairs as the presence of obstacles and soda cans, because states of Herbert's sensors are regularly and reliably caused by those states of affairs.

And we can ascribe contents to representations by drawing on descriptions of Herbert's environment that relate to the goals we ascribe to Herbert. However, our willingness to take these representations seriously as natural representations that bear content intrinsically covaries with our willingness to take Herbert seriously as an organism. We are not as comfortable attributing genuine goals and behaviors to Herbert as we are attributing goals and behaviors to frogs.[2]

Finally, absurd cases like the firing pin can be excluded (for the most part) since guns are not easily construed as "organisms." Firearms are difficult to anthropomorphize, since they do not exhibit autonomous behavioral dynamics and we don't normally see them as having goals of their own. It is not *impossible* to ascribe goals to weapons or other tools, but the ascription of folk-psychological properties to tools, like the folk ascription of a bloodthirsty disposition to a sword, generally depends on the way a tool influences its users' behavior. (I suspect this dependence might offer some novel explanations of why Clark and Chalmers' [1998] extended cognition hypothesis is attractive to some.) The attribution of autonomous behaviors to tools like swords is fanciful. Perhaps we might imagine a tool exhibits psychic "behavior," but anyway we do not believe that swords possess mechanisms that produce this "behavior" (though if we did, such a construal would be more compelling). If the firing pin of a gun is not a component of a behavioral mechanism, it cannot represent anything according to the Organism-Receptor account.

So the Organism-Receptor account licenses an ascriptive practice that resembles the crude receptor notion when the role of construals is not made explicit. It is unusual in that it inverts Ramsey's preferred order of ascription: Ramsey wishes to

---

[2] Notably, Rodney Brooks himself does not claim that it is proper to ascribe representational capacities to Herbert (Brooks, Connell, and Ning 1988; Brooks 1991), but Brooks plausibly had in mind a more demanding account of representation.

ascribe cognitive structure to systems in virtue of their representational structure (see e.g. Ramsey, 222–235), whereas I suggest that we in fact ascribe representational structure in virtue of seeing a system as a system with goal-directed behavior, i.e. as a potentially cognitive system.

6. Worries. Since the Organism-Receptor account shares a certain teleological character with Dretske's account, I will discuss Ramsey's two most developed objections to Dretske, along with other worries specific to the Organism-Receptor account. First, Ramsey objects that Dretske's account is question-begging with regard to the job-description challenge. Roughly, teleological normativity (i.e. functioning and malfunctioning) is not sufficient to explain intentional normativity (i.e. representation and misrepresentation), and since Dretske provides no satisfying criteria for what it is for a state to function as a representation, he cannot bridge that gap (Ramsey 2007, 131–2). But the Organism-Receptor account has more resources than Dretske's teleofunctionalism. Construing a system as an organism involves construing it as exhibiting behavior, which allows us to distinguish behavioral mechanisms from other mechanisms. On the Organism-Receptor account, misrepresentations are malfunctions of behavioral mechanisms (like frog vision), but not of other mechanisms (like a frog's circulatory system or a gun's firing mechanism).

My reply invites a rejoinder: on the Organism-Receptor account the functional roles of representations will be extremely diverse, and representations will be common. They will not just include IO-representation and S-representation (roughly, information-processing relata and models for surrogative reasoning; Ramsey 2007, 68ff.), which Ramsey and most cognitive scientists regard as genuinely representational. They will also include more controversial varieties of "representation," such as Millikan's (1995) "pushmi-pullyu" representations: Janus-faced mechanistic components that simultaneously indicate a state of affairs and cause

an adaptive or designed response. In other words, representations will include what Ramsey calls "causal relays" like the firing pin in a gun, the inclusion of which in the extension of REPRESENTATION was the ground for his reductio! However, the absurd cases can be avoided. The firing pin case is excluded because guns are poor examples of organisms. And pushmi-pullyu representations include cases with significant intuitive appeal to many scientists, like the predator calls of vervet monkeys (Millikan 1995; cf. Seyfarth, Cheney, and Marler 1980). While this conception of representation has a more liberal extension than Ramsey is comfortable with, it is liberal enough to explain common representation-ascriptions in cognitive science without being so liberal as to countenance absurd cases like Ramsey's firing pin, so I submit it is adequate to scientific practice.

Ramsey's second objection is that Dretske is committed to a false principle: that if a component is incorporated into a mechanism because it carries information, then its function is to carry information (132–9). However, the Organism-Receptor account constrains the causal dependence criterion (R1) by relying on construals of systems as organisms instead of teleofunctional commitments. The account I describe is not committed to Dretske's principle, and therefore is not subject to this objection.[3]

Nevertheless, one might worry whether the organism criterion (R2) is a suitable condition on representation-ascription. I suggested five conditions (O1)–(O5) on what can be seen as an organism, but conditions (O1) and (O2) are fairly unconstrained. There are psychological limitations on when goals or behaviors can be plausibly attributed to a system, but what are those limits? And what factors influence interpersonal variability in willingness to make these attributions? The reason this practice isn't bonkers is that it coheres with the explanatory purpose of

[3] Ramsey's discussion is rich and worthy of deeper engagement than this, but for reasons of space I leave the matter here.

representation-ascriptions: to make explicit the quasi-intentional character of behavioral mechanisms. Nevertheless, we should hope that these psychological limitations are vindicated by more principled considerations. Criticism is warranted if scientists attribute goals and behaviors when they should not. There is some extant work on the proper norms ascribing goals to organisms (e.g. Shea 2013; Piccinini 2015, chap. 6), but little serious work on how to understand the concept of BEHAVIOR in the context of cognitive science. We should worry about the practice of ascribing natural representations if scientists construe things that are not cognitive systems as "organisms." Indeed, we might indeed worry that many cognitive scientists misuse the concept COGNITION, given the intense disagreements over its extension (see e.g. Akagi 2017). However, my present aim is not to evaluate scientific practice, but to describe it faithfully (with the hope that a more satisfactory evaluation will follow).

Another worry about construal-based accounts is that they entail an unattractive anti-realism: if representations and their contents only exist relative to construals, they are mind-dependent rather than objective, right? This worry is unfounded. I am undertaking a modified version of Ramsey's job description challenge: my aim is to describe the ascription of representations in virtue of which they serve an explanatory purpose, not to distinguish genuinely representational states from non-representational states. The Organism-Receptor account does not entail that representations exist relative to construals, only that they are *ascribed* relative to construals. My account is consistent with the existence of a first-order account of the metaphysics of representation that justifies this practice (or doesn't). After all, the duck-rabbit can be construed as a duck even if it is not a duck, and nothing about that fact entails that ducks (or unambiguous images of ducks) are not real. The Organism-Receptor account describes a norm that plausibly guides human scientists with imperfect capacities for knowledge. But while my solution to the metadiscursive job description challenge is not inconsistent with Ramsey's solution to

the first-order job description challenge, it is inconsistent with Ramsey's characterization of scientific norms for ascribing natural representations.

7. Conclusion. I began by observing the common worry that scientists ascribe representations more liberally than many philosophers are comfortable with, and in particular that scientists rely on an unsatisfactory "receptor" criterion. I sketched an account on which scientists ascribe natural representations only to components of mechanisms of systems construed as "organisms." Since in practice cognitive scientists attend almost exclusively to systems that are easily so construed, their behavior may appear to be guided by the crude receptor criterion whereas in fact it is guided by the Organism-Receptor criterion. However, while the Organism-Receptor account is still relatively liberal, a crucial difference between the two accounts is that the crude criterion has absurd consequences, whereas such consequences are eliminated or marginalized on the Organism-Receptor criterion. Since scientists do not in fact endorse these absurd consequences, I argue that the augmented criterion is a better hypothesis regarding norms for representation-ascription in cognitive science.

This is proposal is not a comprehensive, new theory of representation, but it accomplishes two things. First, it provides argumentative resources for resisting the common worry that cognitive scientists use hopelessly liberal criteria for ascribing representations. Second, it offers a novel picture of practices for representation-ascription in the biological and behavioral sciences, one that is less pessimistic picture than Ramsey regarding conceptual rigor in cognitive science. The picture is not beyond criticism—in particular, it wants for a more detailed account of the grounds that warrant attributing behaviors and goals to systems. But since it is more faithful to our practice than Ramsey's it is likely to yield more productive suggestions for how to guide that practice into the future. I suggest that we safeguard conceptual rigor in cognitive science not by cleaving more faithfully to the representationalism of the

cognitive revolution, but by embracing role of construal in scientific inquiry, making it explicit, and subjecting it to reasoned criticism.

REFERENCES

Adams, Fred, and Ken Aizawa. 2001. "The Bounds of Cognition." *Philosophical Psychology* 14:43–64.

Adams, Fred, and Rebecca Garrison. 2013. "The Mark of the Cognitive." *Minds and Machines* 23:339–52.

Akagi, Mikio. 2017. "Rethinking the Problem of Cognition." *Synthese*. doi: 10.1007/s11229-017-1383-2.

Bechtel, William, and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421–41.

Brooks, Rodney. 1991. "Intelligence without Representation." *Artificial Intelligence* 47:139–59.

Brooks, Rodney, Jonathan Connell, and Peter Ning. 1988. "Herbert: A Second Generation Mobile Robot." *A.I. Memos* 1016:0–10.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58:7–19.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Cummins, Robert, and Pierre Poirier. 2004. "Representation and Indication." In *Representation in Mind: New Approaches to Mental Representation*, E sds. Hugh Clapin, Phillip Staines and Peter Slezak, 21–40. Amsterdam: Elsevier.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* 60:685–700.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

———. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT/Bradford.

Giere, Ronald N. 2006. *Scientific Perspectivism*. Chicago: University of Chicago Press.

Hubel, David H., and Torsten N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160:106–54.

Jastrow, Joseph. 1899. "The Mind's Eye." *Popular Science Monthly* 54:299–312.

MacFarlane, John. 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Clarendon.

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67:1–25.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

———. 1995. "Pushmi-Pullyu Representations." *Philosophical Perspectives* 9:185–200.

Piccinini, Gualtiero. 2015. *Physical Computation: A Mechanist Account*. Oxford: Oxford University Press.

Ramsey, William M. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.

Roberts, Robert C. 1988. "What Emotion Is: A Sketch." *Philosophical Review* 97:183–209.

Rowlands, Mark. 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: MIT Press.

Rupert, Robert. 2009. *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.

Seyfarth, Robert M., Dorothy L. Cheney, and Peter Marler. 1980. "Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication." *Science* 210:801–3.

Shea, Nicholas. 2013. "Naturalising Representational Content." *Philosophy Compass* 8:496–509.

Stinson, Catherine. 2016. "Mechanisms in Psychology: Ripping Nature at Its Seams." *Synthese* 193:1585–614.

Weiskopf, Daniel A. 2011. "Models and Mechanisms in Psychological Explanation." *Synthese* 183:313–38.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. 3rd Ed. Trans. G.E.M. Anscombe. Eds. G.E.M. Anscombe and Rush Rhees. Oxford: Blackwell, 2001.

Comparing Systems Without Single Language Privileging

Max Bialek

mbialek@rutgers.edu

For the 2018 PSA Meeting.

Word count: 4753

Abstract

It is a standard feature of the BSA and its variants that systematizations of the world competing to be the best must be expressed in the same language. This paper argues that such single language privileging is problematic because (1) it enhances the objection that the BSA is insufficiently objective, and (2) it breaks the parallel between the BSA and scientific practice by not letting laws and basic kinds be identified/discovered together. A solution to these problems and the ones that prompt single language privileging is proposed in the form of privileging the best system competition(s).

1 Introduction

According to the Best Systems Analysis (BSA), the laws of nature are the theorems of the best systematization of the world—with 'best' standardly understood to mean the simplest and most informative (on balance). It is currently a standard feature of the BSA (since Lewis 1983) and its variants (Loewer 2007; Schrenk 2008; Cohen and Callender 2009) that a single language must be privileged as the language in which all systems competing to be the best will be expressed. Two problems have led these authors to adopt single language privileging: The first is the Trivial Systems Problem (TSP), according to which, in brief, allowing for suitably gerrymandered languages can guarantee that the "best" system will have axioms and theorems undeserving of the name "law" (see Lewis 1983 for its initial development). Language privileging provides a quick fix to the TSP as long as the privileged language is not among the suitably (and problematically) gerrymandered. The second is the Problem of Immanent Comparisons (PIC) suggested by Cohen and Callender (2009). The PIC takes it to be the case that there are only "immanent" measures for simplicity, strength, and their balance—that is, measures defined for only one language. With single language privileging, no two systems ever need to be compared when expressed in different languages, and so having to use only immanent measures is not an issue.

Though single language privileging solves these problems for the BSA and its variants, it creates new ones of its own. For one, the BSA is already often criticized for being insufficiently objective—because it is unclear that there is an objective answer to the question of what makes a system the best—and single language privileging has the potential to fuel those criticisms by requiring proponents of the BSA to say which

language gets privileged. Relativizing laws to languages (as in Schrenk 2008 and Cohen and Callender 2009) goes some way to resist such criticisms, but, as Bialek (2017) argues, relativity itself should be minimized (as much as scientific practice allows) when responding to those who employ the 'insufficiently objective' critique of the BSA. Another issue with language privileging—a version of which is suggested in a specific critique of Lewis (1983) by van Fraassen (1989), and is here newly generalized as an issue for *any* single language privinleging—is that it breaks the supposedly close connection in scientific practice between the discovery of the laws and the discovery of basic kinds.[1]

Both problems are, ultimately, overstated, and may be resolved not with single language privileging, but with the privileging of *classes* of languages. This addresses both of the issues just raised. For one, it restores the co-discovery of laws and basic kinds to the BSA by making the search for laws (via a best system competition conducted in the course of scientific practice) include a search through a class of languages for the one that yields the best system-language pair. It also helps to limit the degree to which laws may need to be relativized to language by reducing the problem of privileging a language (class) to the already present problem of choosing a measure of 'best'.

The outline of this paper is as follows. I begin, in Section 2, by laying out the PIC. In Section 3, I argue that the PIC ignores the existence of measures (illustrated by the

[1]Depending on the specific interests of the author, there has been talk of "basic kinds" (as in Cohen and Callender 2009), "fundamental kinds" (Loewer 2007), and "perfectly natural predicates" (Lewis 1983). These are progressively more restrictive ways of interpreting the predicates of a language that appear in the axioms of a best system expressed in that language. Throughout the paper I use the more general phrase "basic kinds", but nothing about that usage precludes a more restrictive reading.

Akaike Information Criterion) that, while not transcendent (since they cannot compare systems expressed in *any* two languages), are also not immanent (since they can compare systems expressed in *some* different languages). Being sensitive to the existence of such measures suggests a slightly different problem of *transcendent* measures, which may be resolved through privileging classes of languages. The problem for single language privileging of breaking the connection between the discovering laws and basic kinds is developed in Section 4, and its resolution via language-class privileging is demonstrated. In Section 5, I argue that the question of which language class to privilege is reducible to the question of which measure(s) of 'best' (simplicity, informativeness, etc.) should be used. Lastly, in Section 6, I note that the reducibility just introduced suggests a new solution to the TSP that is focused on choosing appropriate measures of 'best', with the conclusion being that none of the problems that have prompted language privileging actually require it for their resolution.

2 The Problem of Immanent Comparisons

The "Problem of Immanent Comparisons" (PIC) begins with an appeal in Cohen and Callender (2009) to a distinction in Quine between *immanent* and *transcendent* notions. Quine writes: "A notion is immanent when defined for a particular language; transcendent when directed to languages generally" (Quine 1970, p. 19). Measurements of simplicity, since they depend on the language in which a system is expressed, are taken by Cohen and Callender to be immanent in this Quinean sense. Strength, or informativeness, is similarly immanent, since it is assumed to depend on the expressive power of the language in which a system is expressed. And, to finish out the set, balance

is said to be immanent as well, since it will be a measure dependent on immanent measures of simplicity and strength. If two systems are competing to be the best and are expressed in different languages, then we would need transcendent measures of simplicity, strength, and balance, in order to implement the best system competition. But "there are too few (viz. no) transcendent measures" of simplicity, strength, and balance (Cohen and Callender 2009, p. 8). Cohen and Callender write that

> *Prima facie*, the realization that simplicity, strength, and balance are immanent rather than transcendent—what we'll call *the problem of immanent comparisons*—is a devastating blow to the [BSA and its variants]. For what counts as a law according to that view depends on what is a Best System; but the immanence of simplicity and strength undercut the possibility of intersystem comparisons, and therefore the very idea of something's being a Best System.
>
> (Cohen and Callender 2009, p. 6, emphasis in original)

The only solution to the PIC, since (supposedly) systems can only be compared when they are expressed in the same language, is to adopt single language privileging.

3 Neither Immanent nor Transcendent

The issue with the PIC is that it ignores the existence of a large middle ground of measures that are neither immanent nor transcendent. To start, let us examine the central claim of the PIC: that simplicity, strength, and balance must be immanent measures. In defense of the idea that simplicity is immanent, Cohen and Callender

(2009, p. 5) defer to Goodman (1954) by way of Loewer, who writes: "Simplicity, being partly syntactical, is sensitive to the language in which a theory is formulated" (Loewer 1996, p. 109). Loewer and Goodman are exactly right. Simplicity is language sensitive. For example, let us adopt a naive version of simplicity, $SimpC(-)$, that is measured by the number of characters it takes to express a sentence (including spaces and punctuation). Consider the following sentence.

This sentence is simple.

Its $SimpC$-simplicity is 24 characters. The same sentence in Dutch is

Deze zin is eenvoudig.

The sentence's $SimpC$-simplicity now is 22 characters. So the $SimpC$-simplicity of a sentence depends or is sensitive to the language in which the sentence is expressed. Does that language sensitivity mean that $SimpC$ is immanent? It depends on what is meant by being "defined for a particular language".

$SimpC$ is, in some sense, "defined for a particular language". Insofar as the measure gives conflicting results for a sentence expressed in different languages, it would be ill-defined if we took it to be directed at sentences irrespective of the language in which they are expressed. One way of dealing with this would be to think that we have a multitude of distinct simplicity measures: $SimpC_{\text{English}}(-)$, $SimpC_{\text{Dutch}}(-)$, and so on. But doing that disguises an important fact: each of these measures of simplicity is *the same measure*, just relativized to particular languages. Drawing our inspiration from the "package deal" of Loewer (2007)—in which the BSA holds its competition between system-language pairs (or packages)—we could just as easily deal with the language

sensitivity of *SimpC* by saying it is defined for sentence-language pairs. We don't need, then, different measures of simplicity. Just the one will do:

$$\begin{aligned} SimpC(\ulcorner\text{This sentence is simple.}\urcorner, \text{English}) &= 24 \text{ char.} \\ SimpC(\ulcorner\text{This sentence is simple.}\urcorner, \text{Dutch}) &= 22 \text{ char.} \end{aligned}$$

In this way, *SimpC* is better understood as transcendent, and not immanent, because it is, as Quine put it, "directed to languages generally".

Of course, *SimpC* can't be directed to *all* languages, since it will be undefined for any languages that don't have a written form with discrete characters. This suggest that there is an important middle ground between immanent and transcendent measures. When a measure falls in that middle, as *SimpC* seems to, I will say that it is a "moderate measure".

So which conception of *SimpC* is the right one? The "devastating blow" that immanence deals to the BSA and its variants is that it "undercut[s] the possibility of intersystem comparisons" (Cohen and Callender 2009, p. 6). In our naive example,

$$SimpC_{\text{English}}(\ulcorner\text{This sentence is simple.}\urcorner)$$

is—if *SimpC* is immanent—incomparable to

$$SimpC_{\text{Dutch}}(\ulcorner\text{This sentence is simple.}\urcorner).$$

But obviously it's not. ⌜This sentence is simple.⌝ is *SimpC*-simpler in Dutch than in English (when being *SimpC*-simpler means having a lower value of *SimpC*).

Nothing prevents a transcendent or moderate measure from taking a language as one of its arguments. Such a measure is transcendent (or moderate), but language sensitive, and, importantly, it allows for comparisons even when a variety of languages are involved. That being the case, the mere language sensitivity of simplicity, strength, and their balance is not enough to guarantee that they are immanent, nor is it enough to guarantee the incomparability of systems expressed in different languages.

In response to the existence of a measure like *SimpC*, it might be suggested that there may well be transcendent (or moderate) measures plausibly named "simplicity" (etc.), but these are not the ones relevant to the BSA; the measures that *do* appear in BSA will be immanent. It is absolutely right to question the plausibility of a measure as naive as *SimpC* having a role to play in the BSA. (I certainly do not intend to defend *SimpC* as the right measure of simplicity for the BSA.) But I do not think it is clear why we should assume that the right measures are immanent. Rather, I think that moderate measures are, if anything, the norm, and an example may be found in the selection of statistical models.

Following Forster and Sober (1994), statistical model selection has standardly been associated in philosophy with the Akaike Information Criterion (AIC):

$$AIC(M) \;=\; 2[\text{number of parameters of } M] - 2[\text{maximum log-likelihood of } M]$$

The full details of AIC are not terribly important for our purposes here; it is enough to point out that that first term is concerned with the *number of parameters* of the statistical model M. Forster and Sober note that the number of parameters "is not a merely linguistic feature" of models Forster and Sober (1994, p. 9, fn. 13). But the

number of parameters is *a* linguistic feature of a model. Since AIC can compare models with different numbers of parameters, it can—if we think of statistical models as the system-language pairs of the BSA, and AIC as central to the best system competition[2]—compare systems expressed in different languages. AIC is thus a moderate measure.

It is important to note, however, that AIC is also not a transcendent measure. Kieseppä (2001) offers a response to critics of AIC who are concerned that the measure is sensitive to changing the number of parameters of a model by changing the model's linguistic representation. The response turns on the justification of "Rule-AIC", which says to pick the model with the smallest value of AIC, on the grounds that the predictive accuracy of model M is approximately the expected value of the maximum log-likelihood of M minus the number of parameters of M. Crucially,

> the theoretical justification of using (Rule-AIC) is valid when the considered models are such that the approximation [just mentioned] is a good one.
>
> (Kieseppä 2001, p. 775)

Let M be parameterized to have either k or k' parameters. Then there are two claims that are relevant to the justification of Rule-AIC:

$$\text{predictive accuracy of } M \approx E[(\text{maximum log-likelihood of } M) - k]$$

$$\text{predictive accuracy of } M \approx E[(\text{maximum log-likelihood of } M) - k']$$

[2]To make the connection between AIC and the BSA even stronger, it it worth noting that Forster and Sober (1994) take the "number of parameters" term to be tracking the simplicity of a model.

The predictive accuracy of M is independent of the number of parameters used to express M.[3] But the right side of the approximation in each claim *does* depend on the number of parameters. In general, both of these claims will not be true. Since Rule-AIC is only justified by the truth of these approximations, it will only be applicable to whichever parameterization of M makes the approximation true. The only time when both claims are true, and thus when AIC is applicable to both parameterizations, is when the difference between $E[(\text{maximum log-likelihood of } M) - k]$ and $E[(\text{maximum log-likelihood of } M) - k']$ is negligible. Kieseppä concludes:

> This simple argument shows once and for all that the fact that the number of the parameters of a model can be changed with a reparameterisation does not in any interesting sense make the results yielded by (Rule-AIC) dependent on the linguistic representation of the considered models.
>
> (Kieseppä 2001, p. 776)

From the epistemic perspective that is Kieseppä's concern, I can find room to agree that there is no "interesting sense" in which Rule-AIC is language dependent. This is because, if we are looking to employ Rule-AIC in statistical model selection, what is available to us is a procedure to check if the given parameterization is one that can support the justification of Rule-AIC. If the justification will work, then Rule-AIC applies, and if not, not. Rule-AIC isn't language dependent "in any interesting sense" insofar as it simply doesn't apply to the problematic languages/parameterizations that undermine its justification.

[3]This is intuitively true. It is also true in the formal definition of predictive accuracy given in Kieseppä (1997) and used in this argument from Kieseppä (2001).

However, from the perspective of the BSA and the PIC, these failures of Rule-AIC *are* interesting. AIC (the measure) is not immanent, but it is also not transcendent; it is merely moderate. *Some* reparameterizations of considered models will lead to the inapplicability of Rule-AIC. If Rule-AIC was how we were deciding which system was best, the existence of these problematic reparamterizations would be, as Cohen and Callender put it, a *prima facie* devastating blow to the BSA.

Towards the end of their introducing the PIC, Cohen and Callender write that

> What is needed to solve the problem is a *transcendent* simplicity/strength/balance comparison of each axiomatization against others. The problem is not that there are too many immanent measures and nothing to choose between them, but that there are too few (viz., no) transcendent measures.
>
> (Cohen and Callender 2009, p. 8, emphasis in original)

Cohen and Callender are probably right that there are "too few (viz., no) transcendent measures". In response to this, PIC says that measuring the goodness of a system must be done with immanent measures, and so no systems expressed in different languages may be compared in the best system competition. But non-transcendence is not a guarantee of immanence. We might call the problem that remains the *problem of transcendent measures* (PTC). Measures like AIC are not immanent, but they also aren't transcendent. That non-transcendence gives rise to a degree of language sensitivity that will *sometimes* prevent us from comparing systems expressed in different languages.

In response to the PIC and the supposed immanence of measures appropriate for the BSA, Cohen and Callender (2009) proposed the Better Best Systems Analysis (BBSA),

which relativizes laws to single languages. According to the BBSA, a best system competition is run for every language L (with some restrictions on "every" that aren't especially important here) where all the competing systems are expressed in L and the theorems of the system that is the victor of the competition are the laws *relative to* L. But now it seems that we might have at our and the BSA's disposal moderate measures. In the face of the non-transcendence of these measures—that is, in the face of the PTC—the BBSA's strategy of language relativity is still a good one.[4] Our language relativity does not, however, have to involve privileging *single* languages. The alternative is to relativize to *classes* of languages constructed to ensure the applicability of the measures employed in our best system competition.

4 Discovering Laws and Kinds Together

Before saying more about what relativizing laws to classes of languages would be like in any detail, it is important to say something about why we should pursue language-class relativity over the single language relativity of the BBSA. So, why should we? The reason is that one of the great virtues of the BSA and its variants is their offering of a metaphysics for laws that parallels the search for laws that is to be found in scientific practice, and that parallel is broken by single language privileging. A feature of the

[4]Without going into excessive detail about benefits (and costs) of the BBSA's relativity strategy over competitors, I hope it is enough to note that relativizing the laws allows us to sidestep the question of which language should be privileged entirely, since, ultimately, all languages will get a turn at being privileged, and thus, effectively, none are privileged over all.

search for laws in scientific practice is that it happens in conjunction with a search for the basic kinds of the world. This feature encourages us to acknowledge the importance of language in the BSA, since the basic kinds of the world are, presumably, going to correspond with the basic kinds that appear in the language in which the laws are expressed. Thus, when Lewis first recognizes the language sensitivity of simplicity, he concludes on a celebratory note by saying that the variant of single language privileging he introduces has the virtue of "explaining" why "laws and natural properties get discovered together" (Lewis 1983, p. 368).

For Loewer's Package Deal Analysis, the idea that laws and kinds are discovered together is central to the view. Indeed, the phrase "package deal" has its roots in Lewis, who says just before the "discovered together" remark that "the scientific investigation of laws and of natural properties is a package deal" (Lewis 1983, p. 368). While Loewer ultimately endorses a version of single language privileging, it is accompanied with a rough account of how a "final theory"—i.e., a candidate system-language pair—is arrived at:

> a final theory is evaluated with respect to, among the other virtues, the extent to which it is informative and explanatory about truths of scientific interest as formulated in [the present language of science] SL or any language $SL+$ that may succeed SL in the rational development of the sciences. By 'rational development' I mean developments that are considered within the scientific community to increase the simplicity, coherence, informativeness, explanatoriness, and other scientific virtues of a theory.
>
> (Loewer 2007, p. 325)

If the practice of science parallels the Package Deal Analysis, then the processes of discovering the laws and basic kinds are one and the same.

And it seems Cohen and Callender are also on board with laws and kinds being discovered together when they offer this nice remark on the phenomenon:

> historical disputes between theorists favoring very different choices of kinds seem to us to be disputes between two different sets of laws [...] it has happened in the history of science that people have objected to particular carvings—most famously, consider the outrage inspired by Newton's category of gravity. But given the link between laws and kinds, this outrage is probably best seen as an expression of the view that another System is Best, one without the offending category. If that other system doesn't in fact fare so well in the best system competition—as in the case of the systems proposed by Newton's foes—then the predictive strength and explanatory power of a putative Best System typically will win people over to the categorization employed. While it's true that some choices of [kinds] may strike us as odd, no one would accuse science—the enterprise that gives us entropy, dark energy, and charm—as conforming to pre-theoretic intuitions about the natural kinds of the world. Yet these odd kinds are all embedded in systematizations that would produce what we would consider laws.
>
> (Cohen and Callender 2009, pp. 17–18)

With everyone in agreement, what is the problem? Language privileging, essentially, happens *before* the identification (in the BSA and its variants) or discovery (in scientific practice) of the laws. Though Cohen and Callender will not "accuse science" of

"conforming to pre-theoretic intuitions about the natural kinds of the world", that is exactly what the BBSA (and any other single language privileging variant of the BSA) does when it privileges sets of kinds prior to a best system competition. Furthermore, PIC makes it such that "the predictive strength and explanatory power of a putative Best System" cannot "win people over to the categorization employed" because comparing two putative Best Systems expressed in different languages (with different "categorizations") is supposed to be impossible.[5]

Relativizing to classes of languages solves this problem. Scientists are able to approach the discovery of laws and kinds with pre-theoretic intuitions about how to systematize the world, the language to use when doing that, and the best system competition. As we will see below, the intuitions regarding language and the best system competition will locate them in a particular language class. Scientists will move away from their intuitions about language (and systematizing) when, much as Loewer describes above, there are languages in the relevant language class that may be paired with systems to yield a system-language pair that is scored better by the best system competition than the pre-theoretic system-language pair.[6]

[5]At least, it is impossible according to PIC for the BSA and its variants. If it *is* possible for scientists, then it is wholly unclear why it would be impossible for the BSA.

[6]This movement is only metaphorical for the BSA, where all the possibilities are considered and judged simultaneously. It is helpful, though, to think in the more methodical terms—of considering particular transitions from one system-language pair to another, the benefits that they might bring, and then adopting them or not—because that is what will happen in actual scientific practice.

5 Limiting Language Relativity

Let us begin addressing how language-class relativity can work by looking in more detail at the single language relativity of the BBSA. In the BBSA, there are the fundamental kinds K_{fund}. The set of all kinds $\mathcal{K}$ is the set including K_{fund} closed with respect to supervenience relations—that is, $\mathcal{K}$ includes every kind that can be defined as supervening on the arrangement of the K_{fund} kinds in the actual world. A language L is determined by the set of kinds for which it has basic predicates, and there is a language L_i for every $K_i \subseteq \mathcal{K}$. For any two languages L and L', the supervenience relations between the kinds of the languages and K_{fund} can be thought of as schemes for *translation* between L and L'. The set of all languages $\mathcal{L}_{\text{all}}$ can be thought of as the set of languages that includes L_{fund} closed with respect to all translations. A class of languages $\mathcal{L}_i$ is a set of languages including L_{fund} closed with respect to some acceptable (all, in the case of $\mathcal{L}_{\text{all}}$) translations.

To illustrate, let us consider a 'coin flip' world. Such a world is a string of `H`s and `T`s, which we will assume are the only two fundamental kinds. Another set of kinds might be $K_{\text{ex}} = \{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}\}$, where the translation that gets us to the corresponding language L_{ex} from L_{fund} maps the pairs `HH`, `HT`, `TH`, and `TT`, to `a` through `d`, respectively. An example of a class of languages that includes L_{ex} could be $\mathcal{L}_{\text{n-tuple}}$: Let an acceptable translation for $\mathcal{L}_{n\text{-tuple}}$ be one that, for a given n takes the set of all n-tuples of `H` and `T`, and maps them to a set of kinds $K_n = \{k_{n,1}, k_{n,2}, ...k_{n,2^n}\}$. L_{fund}, then, is just L_1. When `a` through `d` are $k_{2,1}$ through $k_{2,4}$, our K_{ex} and L_{ex} are precisely K_2 and L_2. All, and only, the languages that may be formed through this procedure will be members of the class $\mathcal{L}_{\text{n-tuple}}$.

A language-class relative variant of the BSA will run a best system competition for

every class of languages $\mathcal{L}_i$. Then $\mathcal{S}$ is the set of all systematizations of the world, the set of all competing system-language pairs for the $\mathcal{L}_i$-relative best system competition is given by $\mathcal{S} \times \mathcal{L}_i$.

We can apply this conception of language-class relativity to our other running example of statistical model selection with AIC. Recall that *some* reparameterizations of statistical models would prove problematic for the use of AIC. To reparameterize a model is akin to translating it from one language to another. We can understand, then, the problem of language sensitivity for AIC as being related to some set of problematic translations. If we subtract these problematic translations from the set of all translations, then we have a set of acceptable translations which defines a class of languages that we can call $\mathcal{L}_{AIC}$. $\mathcal{L}_{AIC}$ is precisely the set of all languages such that a system expressed in any one of them will be comparable to a system expressed in any other using AIC. As long as the moderate measures used in the best system competition have clearly problematic and/or acceptable translations associated with them, then the class of languages that may be used to express competing systems will be determined by the measures used in the best system competition.

This will have one of two effects on the extent to which the BSA must be relativized to classes of languages, but before going into those details it will be helpful to characterize "competition relativity". Competition relativity should be understood in much the same way that language relativity is understood. The competition of the BSA is the thing that takes system-language pairs as its inputs, and outputs a best pair from which we can read off the laws. The competition decides what system-language pair is best by considering how well they measure up with respect to some collection of theoretical virtues (like simplicity and informativeness) and the actual world. Much as

we might worry about what language to privilege, and side-step that problem by relativizing laws to languages so that every language takes a turn as the privileged one, we might also worry about which competition, or which set of theoretical virtues, to privilege. Competition relativity sidesteps the problem of which collection of theoretical virtues to use (and weighting between them, and means of measuring them, etc.) by relativizing laws to every way of formulating a best system competition.[7]

So, either the BSA will be committed to competition relativity or not. Suppose that it is not. For convenience, suppose further that Rule-AIC is all that there is to the best system competition. In that case, the BSA will always be run using the $\mathcal{L}_{AIC}$ class of languages. Language-class relativity is not required since there is only one language class that will ever be relevant to the BSA—namely $\mathcal{L}_{AIC}$, as determined by the best system competition. Now suppose that there is competition relativity. A different best system competition must be run for every competition function C_i in the set of all possible competition functions $\mathcal{C}$. In principle we will need to run best systems competitions for every pair in $\mathcal{C} \times \mathbb{L}$, where $\mathbb{L}$ is the set of all language classes. Let $\mathcal{L}_j$ be the class of languages constructed according to the translations that are acceptable for the measures that comprise C_i when $i = j$. In practice, however, it will only make sense to run a competition once for each $C_i \in \mathcal{C}$, since the pairs $C_i, \mathcal{L}_j$ will be unproblematic only when $i = j$. Language-class relativity in this situation will be redundant with competition relativity. We also have it that, in either case (of needing competition relativity or not), single language relativity remains unnecessary for all the same reasons that recommended language-class relativity.

[7]See Bialek (2017) for an extended discussion of competition relativity and the possibility of its inclusion in the BSA.

6 The Trivial Systems Problem

The redundancy of any sort of language privileging relativity with competition relativity offers an interesting solution to the Trivial Systems Problem (TSP) that initiated the trend of single language privileging.

Recall that the TSP is concerned with the possibility of suitably gerrymandered languages that can guarantee that the "best" system will have axioms and theorems undeserving of the name "law". In the introduction fo the problem, Lewis imagines a system S and predicate F "that applies to all and only things at worlds where S holds" (Lewis 1983, p. 367). The system S, then, maybe be expressed by the single axiom $\forall xFx$, simultaneously achieving incredible informativeness—because of the specific applicability of F—and incredible simplicity—because, Lewis assumes, '$\forall xFx$' is about as simple as a system could be. So S will be the best system despite a variety of reasons why it shouldn't be, the foremost of which are that: (1) $\forall xFx$ will be a law unlike any we would expect to find, (2) F would be a basic kind unlike any we would expect to find, and (3) every regularity of the world is a theorem of $\forall xFx$, so there would be no distinction between accidental and lawful regularities.

The problem is solved as long as we can avoid languages that include problematic predicates like F. Single language privileging solves this problem as long as the privileged language does not include the (or any) problematic predicate(s). Language-class privileging likewise solves the problem as long as no language in the class includes the (or any) problematic predicate(s). That alone might be enough said, but the redundancy of language-class choice on competition choice offers a more nuanced solution: The best system competition could be chosen such that the corresponding class

of languages does not include F or any similarly problematic predicates. But it could also be chosen such that F and its ilk are certain to not be the best. Lewis assumes with no discussion that $\forall xFx$ is an incredibly informative and simple system, but, even if that is true for the measures/competition, it need not be true for every competition. If there is competition relativity, then there may be competitions for which a trivial system like $\forall xFx$ is the victor, but for the same reasons that such a system is problematic, scientists will simply be uninterested in the laws relative to those competitions.[8] If there isn't competition relativity, it seems unlikely that science would unequivocally endorse a competition that yields a trivial system (or, if it does, then we would need to take a step back and seriously reconsider our aversion to such a system).

In the end, there is no apparent need for any language privileging or relativity in the BSA.[9] Its role in solving the problems of immanent (or transcendent) comparisons and trivial systems will be unnecessary (if a single moderate best system competition can be identified) or redundant with competition relativity.

[8]In much the same way that Cohen and Callender (2009) allow for there to be uninteresting sets of laws determined relative to languages that include F-like predicates.

[9]The problems discussed is not the only reason one might want to adopt language relativity in the BSA. It should also be noted that one of the virtues of the BBSA's single language relativity is that it allows the view to accommodate an egalitarian conception of special science laws. Language relativity, however, is not the only way of getting special science laws out of the BSA. This is an important issue to which the discussion in this paper is relevant, but a proper exploration of it warrants a more focused and extended treatment.

References

Bialek, M. (2017). Interest relativism in the best system analysis of laws. *Synthese 194*(12), 4643–4655.

Cohen, J. and C. Callender (2009). A better best system account of lawhood. *Philosophical Studies 145*(1), 1–34.

Forster, M. and E. Sober (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science 45*(1), 1–35.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Kieseppä, I. (1997). Akaike information criterion, curve-fitting, and the philosophical problem of simplicity. *The British journal for the philosophy of science 48*(1), 21–48.

Kieseppä, I. (2001). Statistical model selection criteria and the philosophical problem of underdetermination. *The British journal for the philosophy of science 52*(4), 761–794.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Loewer, B. (1996). Humean supervenience. *Philosophical Topics 24*(1), 101–127.

Loewer, B. (2007). Laws and natural properties. *Philosophical Topics 35*(1/2), 313–328.

Quine, W. V. O. (1970). *Philosophy of logic*. Harvard University Press.

Schrenk, M. (2008). A theory for special science laws. In S. W. H. Bohse, K. Dreimann (Ed.), *Selected Papers Contributed to the Sections of GAP.6*, pp. 121–131. Paderborn: Mentis.

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.

Explaining Scientific Collaboration: a General Functional Account

Thomas Boyer-Kassem* and Cyrille Imbert†

October, 2018

Abstract

For two centuries, collaborative research has become increasingly widespread. Various explanations of this trend have been proposed. Here, we offer a novel functional explanation of it. It differs from accounts like that of Wray (2002) by the precise socio-epistemic mechanism that grounds the beneficialness of collaboration. Boyer-Kassem and Imbert (2015) show how minor differences in the step-efficiency of collaborative groups can make them much more successful in particular configurations. We investigate this model further, derive robust social patterns concerning the general successfulness of collaborative groups, and argue that these patterns can be used to defend a general functional account.

*MAPP (EA 2626), Univ. Poitiers, France. thomas.boyer.kassem@univ-poitiers.fr
†CNRS, Archives Poincaré, France. cyrille.imbert@univ-lorraine.fr

1 Introduction

For two centuries, co-authoring papers has become increasingly widespread in academia (Price, 1963, Beaver and Rosen, 1979), especially in the last few decades. Since the 1950s, the percentage of co-authored papers has grown at a common rhythm for science and engineering, social sciences, and patents; the mean size of collaborative teams has also increased, and even more so in science and engineering. No such increase is visible for the art and humanities (Wuchty et alii, 2007).

Various explanations of this collaborative trend have been proposed: for example, it may be caused by scientific specialization, it may increase the productivity or reliability of researchers, or be promoted by the rules of credit attribution. Here, we aim at offering a new functional explanation of this trend by showing that collaboration exists because it increases the successfulness of scientists. The present explanation differs from accounts like that of Wray (2002) by the social and epistemic mechanism that grounds the beneficialness of collaboration. We analyze further an existing model that shows how minor differences in the step-efficiency of collaborative groups at passing the steps of a project can make them much more successful in particular configurations (Boyer-Kassem and Imbert, 2015) and show how it can be used to build a general and robust functional explanation of collaboration.

We introduce the model in section 2. After presenting functional explanations (section 3), we show how the model can be used to derive robust social patterns of the successfulness of collaborative groups (section 4), and argue that these patterns can refine and strengthen functional explanations of collaboration like the one defended by Wray (sections 5 and 6).

2 Boyer-Kassem and Imbert's Model: Main Results and Explanatory Lacunas

Boyer-Kassem and Imbert (2015) investigate a model in which n agents struggle over the completion of a research project composed of l sequential steps. At each time interval, agents have independent probabilities p of passing a step. When an agent reaches the end of the project, she wins all the scientific credit and the race stops (this is the priority rule). Agents can organize themselves into collaborative groups for the whole project, meaning that they only share information, i.e. step discoveries — clearly, there are more favorable hypotheses associated with collaborating, like having new ideas or double-checking (see below). Within a group, agents make progress together, and equally share final rewards. Thus, a group of k agents (hereafter k-group) passes a step with probability $p_g(k,p) = 1-(1-p)^k$. In forthcoming illustra-

tions, the value of l is set to 10 and that of p to 0.5, which is not particularly favorable for groups (ibidem, 674). If collaboration is beneficial with these hypotheses, it will be even more so with more favorable or realistic ones. A community of n agents (hereafter, n-community) can be organized in various k-groups. For example, a 3-community can correspond to configurations (1-1-1), (2-1) or (3). The individual successfulness of an agent in a k-group in a particular configuration is defined as the average individual reward divided by time. It has been obtained for all configurations up to $n = 10$, on millions of runs.

Note that this model is not aimed at quantifying the actual successfulness of collaborative agents, but at analyzing the differential successfulness of agents depending on their collaborative behavior. The main finding is that minor differences in the efficiency at passing steps can be much amplified and that, even with not-so-favorable hypotheses, collaboration can be extremely beneficial for scientists. For example, in a (5-4) (resp. (2-1)) configuration, whereas the difference in step efficiency between the 5 (resp. 2) and the 4-group (resp. 1-group) is 3% (resp. 50%), the difference in individual successfulness is 25% (resp. 700%). The scope of these results actually goes beyond the initial hypotheses in terms of information sharing. Formally speaking, the model is a race between (collective) agents i with probabilities p_i of passing steps. *Whatever the origin* of the differences in p_i, they are greatly amplified by the sequential race. In other words, any factor, whether epistemic or not, that implies an increase in p_i of a k-group (e.g. if a collaborator is an expert concerning specific steps, if increased resources improve step-efficiency, etc.) makes this group as successful as a larger group — hence the generality of this mechanism.

Still, these results do not explain scientific collaboration by themselves. First, collaboration is beneficial for particular k-groups in particular configurations only: a 2-group is very successful in configuration (2-1-1-1-1) but not in (7-2). Thus, the model mostly provides possibility results about what can be the case in certain configurations. Second, the explanandum is a general social feature of modern science, not some collaborative behavior in some particular case, so the explanans must also involve general statements about the link between collaboration and beneficialness. Then, if the model presents generic social mechanisms with explanatory import, one needs to describe at a general level the effects of these mechanisms and provide some general, invariant pattern between collaboration and beneficialness. This is what we do in section 4. A final serious worry is that the beneficialness of a state by no means explains why it exists, nor perseveres in being. A link needs to be made between the beneficialness of collaboration and its existence over time. We suggest that this connection can be accounted for functionally.

3 Functional Explanations and Collaboration

We review in this section how functional explanations work and how they can be used in the present case. We follow Wray's choice to use Kincaid's account because it is simple, widely accepted, and that nothing substantial hinges on this choice. Functional explanations explain the existence of a feature by one of its effects, usually its usefulness or beneficialness. As such, they can be sloppy and badly flawed. The usefulness of the nose to carry glasses does not explain that humans have one. Nevertheless, if stringent conditions are met, it is usually considered that functional explanations can be satisfactory, typically within biology. Even Elster, who otherwise favors methodological individualism, agrees that functional explanations can be acceptable in the social science (Elster, 1983). According to Kincaid (1996, 105-114), P is functionally explained by E, i.e. P exists "in order to promote <effect E>" if:

(1) P causes E,

(2) P persists because it causes E,

(3) P is causally prior to E.

Then, a functional explanation of collaboration should have the following form:

(1c) Scientists' collaborative behavior causes the increase of their individual successfulness.

(2c) Scientists' collaborative behavior persists (or develops) because it causes a higher individual successfulness.

(3c) Collaborative behavior is causally prior to this increased individual successfulness that is rooted in collaborative behavior.

We agree with Wray (2002, 161) that it is implausible to consider that the high successfulness of scientists is the initial cause of collaboration since many scientists have been successful (and continue to be in some fields) without collaborating. In the same time, there can be various contingent reasons why some researchers have decided to engage in some collaboration. So, what calls for an explanation is the fact that collaboration is widespread and persistent, not its occasional existence.

4 Collaboration Causes Successfulness

We now argue that the above model provides strong evidence in favor of (1c). To explain the general collaborative patterns described above, the causal

relation between collaboration and successfulness needs to be general and robust. Hence, one needs to go beyond the description of the beneficialness of collaboration in particular situations. A first route is to find general results about when it is beneficial for individuals to collaborate, such as the following theorem (see the appendix for the proof).

Theorem. When m groups of equal size k merge, the individual successfulness of agents increases.

In other words, as soon as several k-groups of the same size exist, they would improve the individual successfulness of their members by merging. A corollary is that single individuals always have interest in collaborating. However, this theorem only covers a small subset of possible configurations, and cannot provide a general vindication for the causality claim (1c). Further, agents might only use it if they are aware of it and are in a position to identify groups of equal-size competitors, which cannot be assumed in general.

To overcome these difficulties, we now assess agents' successfulness irrespective of what they know about other competitors: we consider the average successfulness of k-groups over all possible configurations for each community size. For example, we average the individual successfulness of 4-groups in configurations (4-1-1-1); (4-2-1) and (4-3)[1]. In order to study the robustness of the causal relation between collaboration and successfulness, we investigate in the next paragraphs how much collaborating remains beneficial under variations of key parameters of the competition context.

Successfulness and community size. Figure 1 shows the average successfulness within k-groups for communities of various sizes. First, the successfulness of loners brutally collapses and is much lower than that of other k-groups as soon as $n > 2$. This confirms that except when nobody collaborates, or in very small communities, loners are outraced. Second, for all group sizes, individual successfulness decreases for larger communities, as can be expected when the number of competing groups and their size increases. Nevertheless, the successfulness of k-groups remains high and stable up to some community size s larger than k till they are eventually outperformed by larger groups or till growing bigger would mean over-collaborating (see (Boyer-Kassem and Imbert, 2015, 679-80) for an analysis of over-collaboration in large groups). Third, the larger the groups are, the longer and flatter this initial plate of successfulness is and the less steep the decrease in successfulness is. Fourth,

[1]There is no clear rationale about how to weigh configurations. From a combinatorial viewpoint, configuration (1,1,1,1,1,1,1) has one realization and (3,2,1,1) several ones. But from an empirical viewpoint, when scientists hardly collaborate, configuration (1,1,1,1,1,1,1) is usual and (3,2,1,1) extremely rare. We have privileged simplicity and chosen to give equal weight to all configurations.

when n is much larger than k, the successfulness of k-groups increases with k. However, this increase is a moderate one and small groups still do reasonably well, which is somewhat unexpected, given the general amplification effect — but see the analysis of figure 3 below for more refined analyses. Typically, in 10-communities, 2-groups do badly but remain somewhat viable since their average successfulness remains between 1/3 to 1/2 of that of 3 or 4-groups. Overall, not collaborating is in general not a viable strategy. Collaborating moderately ($k = 2$ or 3) can be very rewarding when there are few competitors (e.g. in small research communities, or on ground-breaking questions that are only known to a handful of scientits). Smalls groups remain viable but tend to be outraced when communities become significantly larger (typically, concerning questions belonging to normal science that many researchers are likely to tackle). Thus, moderately collaborating is a viable but more risky strategy when uncertainty prevails about the number and size of competing groups. Finally, while large collaborative groups rarely get exceptionally high gains, they are extremely safe, with moderate differences in successfulness between them or when the community size increases.

Successfulness and group size. Figure 2 shows the variation of individual successfulness with group size for various community sizes. First, for $n > 2$, the successfulness curve has a one-peaked (discrete) form, the maximum of which grows with the community size. Second, these one-peaked curves are not symmetric: the increase in successfulness is steep (but less so for larger groups), the decrease is gradual (idem). Large groups predate resources so groups need to grow big quickly to get some share and because returns can be increasing (Boyer-Kassem and Imbert 2015, 678), the increase in successfulness is steep. The decrease after the peak is slow because large group are hard to predate but over-collaborating can become suboptimal when the increase in gain by predation no longer makes up for the need to share between more people). These results are not trivial because at the configuration level, the successfulness of groups is contextual. They are important, too. A one-peaked profile is usually *assumed* in the literature about coalitions. Here, it emerges from a micro-model, and gets its justification from it. Overall, these patterns show again that agents have a large incentive to collaborate substantially, whatever the competing environment.

Successfulness in more or less collaborative communities. Figure 3 finally shows how the successfulness of k-group members varies with the degree of collaboration in their competition environment.[2] Here again, what matters

[2]Here, the degree of collaboration in each configuration is assessed by computing the average size of k-groups. For each k, we then compute the average successfulness of a member of a k-group over configurations having a degree of collaboration within intervals $[1, 1.5]$ (represented at coordinate "1.25" on the x-axis), $[1.25, 1.75]$, $[1.5, 2]$ ... $[3.5, 4]$. We

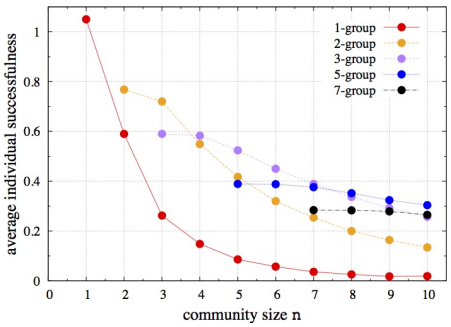


Figure 1: Variation of individual successfulness with community size.

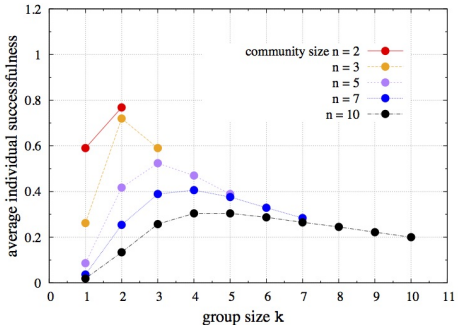


Figure 2: Variation of individual successfulness with the size of groups.

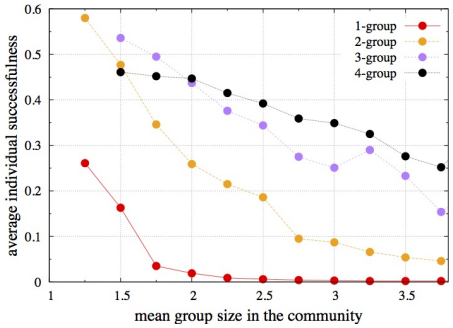


Figure 3: Variation of successfulness with the degree of collaboration in communities.

is less the exact value of the successfulness than the differential successfulness between more or less collaborating individuals. The graph confirms that successfulness depends less on the absolute size of groups than on how much they collaborate in comparison with their competitors. Scientists who collaborate more than average are very successful; those who collaborate as their peers do reasonably well; those that collaborate less than average are outraced by a large margin. This general result is not unexpected given all the above results, but the graph highlights that success for intensively collaborating scientists, and underachievement for under-collaborators can be very large. This is an important finding because if, as we shall see, successful scientists pass over their collaborative habits more than their peers, then the feedback loop provides a mechanism that favors the *increase* of the degree of collaboration by promoting those that collaborate more than others.

have chosen overlapping intervals to smoothen results. The average is computed up to communities of size 10.

Partial conclusion. Overall, the results show that — everything else being equal — collaborating a lot entails successfulness. This relation is robust under changes in the size of communities or in the exact size of groups. Further, those who collaborate more than average are much more successful. Collaborating too much is not a significant problem, under-collaborating is. So, collaborating a lot is a safe working habit, especially in the absence of information about the size and structure of the competing community. In light of this evidence, (1c) seems adequately supported.

5 Collaborative Practices Develop Because of the Success of Collaborative Scientists

We have so far argued that collaborative scientists, especially when they collaborate more than others, are more successful. We now need to argue that, because of this differential successfulness, collaborative habits persist and possibly develop in scientific communities (2c). A wide variety of social mechanisms across scientific contexts can contribute to this feedback loop. Accordingly, we shall be content with giving various evidence that strongly suggests that this link is a likely one.

Transmission. Knowing how and when to collaborate is not straightforward. Like other know-how skills, it can be developed by exercising it with people who already possess the relevant procedural knowledge. In this case, people who already collaborate can endorse this role of cultural transmission for colleagues and above all students (Thagard, 2006). Working with students is an efficient way to train them as scientists (Thagard, 1997, 248—50), so scientists have incentives to enroll students in their collaborative groups. Then, the cultural transmission of collaborative practice does not require any particular effort on top of that. The very circumstances that make collaboration possible and beneficial also make its transmission easier: when a research project can be divided into well-defined tasks, the solutions of which can be publicly assessed and shared, it is easier to enroll other people and thereby transmit collaborative skills to them (ibidem). Thus, collaborative habits can be passed over and need not be reinvented by newcomers.

Transmission opportunities. We now argue that collaborative scientists, because they are more successful, will more often be in a position to transmit their collaborative habits and that the collaboration rate will therefore increase. Within applied science, in which collaboration is also widespread (Wuchty, 2007), research projects are usually directed at finding profitable applications, which can be patented. Thus, fund providers are directly and strongly interested in hiring and providing resource to successful scientists,

who develop such applications. Within pure science, the connection is less straightforward. But because scientific success is the official goal of science, successful scientists can be expected to stand better chances to get good positions and grants, develop research programs, and pass over their collaborative habits.

Note that it is merely needed that the function between the pragmatic rewards of scientists and their success is on average increasing. This remains compatible with the fact that *some* epistemically successful scientists get little resource and *some* unsuccessful scientists get a lot — which seems to be the case. Actually, non-epistemic factors may even tend to over-credit successful scientists, and in particular collaborative ones. First, individual successfulness has been assessed in the model with a conservative estimate. It seems that an agent's publication within a k-group is actually more appreciated than just $1/k$ of a single-authored publication. For instance, a large French research institution in medicine officially weighs the citations of a paper with "a factor 1 for first or last author, 0.5 for second or next to last, and 0.25 for all others" (Inserm 2005). Also, a publication within a 10-group will generally be more visible than one single-authored publication, since more people can promote or publicize collective publications and research topics. Second, sociology of science seems to indicate that scientific credit tends to accrue to a subset of scientists who are perceived as extremely successful — this is the Matthew effect (Merton, 1968). Then, to the extent that access to resources increases with scientific credit, successful collaborative scientists can be expected to benefit from this effect and transmit more their working habits. The concentration of credit and resource may further stimulate collaborative behavior with these fortunate scientists.

Other types of mechanisms may contribute to this process, like conscious ones. So far, agents have only been supposed to follow their working habits and sometimes transmit them. But supplementary intentional or imitative processes may also feed this dynamics [3]. Once winners of the scientific race publish co-authored articles, it becomes easy for others to see that successful scientists are highly collaborative ones. (For instance, if agents of a 3-group are 4 times more successful than a single agent, this means that their groups publishes 12 more articles than this agent). Accordingly, the belief that collaborating is beneficial can be acquired as collaborating becomes usual. Furthermore, resources may accrue to scientific institutions that host individually successful scientists, and indirectly to these scientists. Agents in the model can be reinterpreted as teams or collective entities which decide to share results or to combine their expertise to produce collective articles. Then, these institutions

[3]Kincaid mentions that "complex combinations of intentional action, unintended consequences of intentional action, and differential survival of social practices might likewise make these conditions [(1)–(3) in our Section 3] true" (Kincaid 1996, 112).

and their members will be more successful, may attract resource, and will keep developing and transmitting their working habits.

In light of the above discussion, we believe that the causal connection between the success of collaborative scientists and the persistence and development of collaborative practices is highly plausible.

6 Discussion

Good functional explanations should be unambiguous about when the causal mechanisms that they rely on are efficient. In the present case, the following conditions can be emphasized.

First, conditions for the application of the priority rule should be met. In particular, (i) it should be possible to single out problems and to state uncontroversially when they are solved. Second, for the model to apply, (ii) scientific problems should be dividable into subtasks, and (iii) the solutions of these subtasks should be communicable. Finally, the model assumes that (iv) the completion of these subtasks should be sequential, but our conclusions still hold if this condition is relaxed. Indeed, if some subtasks can be tackled in parallel then the project can be completed even more quickly by different agents of a group, and collaboration is even more successful. Conditions (i)-(iii) are somewhat met in the formal and empirical sciences, less so in the social science, and almost not in the humanities. For example, as noted by Thagard (1997, 249), the humanities do not obviously lend themselves to the division of labor and to teacher/apprentice collaborations. Similarly, the importance of interpretative methods and the coexistence of incompatible traditions may prevent consensus on the nature of significant problems and what counts as a solution. This may account for the differences concerning collaborative patterns in these fields.

As mentioned above, different causal pathways may connect the successfulness of collaborative scientists to the persistence and development of collaborative practices. Thus, conditions for the fulfillment of claim (2c) cannot be uniquely specified. But several points are worth mentioning. First, the activity of epistemically successful scientists should be favored by scientific institutions. This can be the case if it is agreed that scientific success, in the form of publications or patents, is valued and promoted. Concerning scientific results that lead to patents, applications and financial gains, this condition is met when public or private funders value such outputs. Concerning pure scientific results, this means that there should be a wide agreement about which results are scientifically good and significant, and there should exist common and accessible publication venues, the value of which is consensual. Again, these conditions are approximately met in the formal and empirical sciences, less so in the social science and, almost not in the humanities in which scholars do not share paradigms, methods or norms about what is scientifically sound

and significant, and cultural and linguistic barriers can restrain the existence of unified communities and common publication venues. Second, in contexts in which researchers and projects are regularly evaluated, especially by agents or institutions who are not in a position to asses the scientific value of their work, the existence of a common standard of success in terms of publications (through simple and calibrated publication indicators) may even more favor researchers who are successful, and therefore the development of collaboration. Finally, when resources are crucial to carry out or facilitate research, snowball effects can favor even more successful scientists, and in particular collaborative ones. This resource accessibility condition, which is central in Wray's explanation, is not in ours. But we agree that in such cases, the functional mechanisms that we describe will be even stronger. In this sense, our account encompasses Wray's. This condition about resources may be another reason for the difference in collaborative behavior between the formal or empirical sciences, the social sciences and the humanities.

7 Conclusion

We have argued that collaborating a lot is overall a safe and success-conducing practice. This conclusion is robust for various sizes of groups, communities and degrees of collaboration; everything being equal, those who collaborate more than average do better. Then, to the extent that the successfulness of researchers gives them more opportunities to transmit their research habits, the development of collaborative practices in communities can be functionally explained. We have further emphasized that the conditions for this functional pattern to work are specifically met in the scientific fields in which collaboration is well-developed. Accordingly, it seems reasonable to consider that this functional mechanism is an important element of the explanation of the development of collaboration in modern science.

The explanation of collaboration is probably a multi-factorial issue. Nevertheless, an asset of our general functional explanation is that it highlights the unexpected force of beneficial aspects of collaborative activities and suggests important roles for contextual factors that are associated with the rise of collaboration. As such, it is general and unifying. For instance, the competition model shows how the division of scientific labor, the use of specialized experts (Muldoon 2017), or the increased reliability of collaborative teams (Fallis 2006, 200) can increase the probability that groups pass research steps and have amplified effects in terms of successfulness. Similarly, factors like the need to access resources to carry out or facilitate research can create a snowball effect that favors epistemically successful (collaborative) researchers (Wray 2002). And factors like the globalization of research or professionalization (Beaver, 1979) can be seen as conditions favoring the application of the priority rule

and scientific competition.

Finally, while nothing in the model provides an internal limit to the growth of collaboration, one can note that there is a wealth of reasons why collaborating groups cannot develop forever. For example, communities are limited in size, spatially distributed, and collaboration is all the more costly as groups are large. The model could be easily modified to integrate factors that limit the success and development of collaboration.

8 Appendix: Proof of the Theorem

Consider first the simple case where the m k-groups don't have other competitors. By symmetry, all groups have the same probability $1/m$ to win the race and get the reward — call this reward r. So, the individual expected reward is $r/(km)$. Suppose now the groups merge and all km agents collaborate. Each of them will receive the same reward, so their expected individual rewards are $r/(km)$ too. However, what matters in the model is not the expected reward, but the successfulness, which is this quantity divided by time. Because within a collaboration agents share all the steps they pass, the larger km-group will be at least as quick, and sometimes more, than all k-groups — more precisely: for a given drawing of all random variables corresponding to attempts to pass the steps, for all agents and temporal intervals, the km-group will move at least as quickly as all k-groups. So the individual successfulness is at least as high when identical groups merge.

Consider now the case where there are other competitors than the m groups. For a given drawing of all random variables, either the winner is one of the m groups, or another competitor. In the former case, the above reasoning can be made again, and the same conclusion holds. In the latter case, there is nothing to lose, and because the km-group is sometimes quicker than the m k-groups, there can be additional cases where it outcompetes the other competitors; then, the individual successfulness increases with the merging. QED.

9 References

Beaver, Donald deB. and Rosen, Richard (1979) "Studies in Scientific Collaboration: Part III", *Scientometrics*, 1(3): 231-245.

Boyer-Kassem, Thomas, and Cyrille Imbert (2015), "Scientific Collaboration: Do Two Heads Need to Be More than Twice Better than One?" Philosophy of Science 82 (4): 667–88.

Elster, Jon (1983), *Explaining Technical Change: A Case Study in the Philosophy of Science*, Studies in Rationality and Social Change, New York: Cambridge University Press.

Fallis, Don (2006), "The Epistemic Costs and Benefits of Collaboration", *Southern Journal of Philosophy* 44 S: 197–208.

INSERM (2005), "Les indicateurs bibliométriques à l'INSERM", `https://www.eva2.inserm.fr/EVA/jsp/Bibliometrie/Doc/Indicateurs/Indicateurs_bibliometriqu Inserm.pdf`

Kincaid, Harold (1996), *Philosophical Foundations of the Social Sciences*, Cambridge University Press.

Merton, Robert K. (1968), "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered", *Science*, 159 (3810): 56–63.

Muldoon, Ryan (2017), "Diversity, Rationality, and the Division of Cognitive Labor", in Boyer-Kassem, T., Mayo-Wilson, C. and Weisberg, M. (eds.), *Scientific Collaboration and Collective Knowledge*, New York: Oxford University Press.

Price, Derek John de Solla (1963), *Little Science, Big Science*, New York, Columbia University Press.

Thagard, Paul (1997), "Collaborative Knowledge", *Nous* 31(2): 242—261.

— (2006), "How to Collaborate: Procedural Knowledge in the Cooperative Development of Science", *The Southern Journal of Philosophy*, XLIV: 177—196.

Wray, K. Brad (2002), "The Epistemic Significance of Collaborative Research", *Philosophy of Science* 69 (1): 150-168.

Wuchty, Stefan, Jones, Benjamin F. and Uzzi, Brian (2007), "The Increasing Dominance of Teams in Production of Knowledge", *Science* 316(5827): 1036-1039.

Individuating Genes as Types or Individuals: Philosophical Implications on Individuality, Kinds, and Gene Concepts

Ruey-Lin Chen
Department of Philosophy
National Chung Cheng University
This paper will be presented at PSA 2018 meeting at Seattle in November

Abstract

"What is a gene?" is an important philosophical question that has been asked over and over. This paper approaches this question by understanding it as the individuation problem of genes, because it implies the problem of identifying genes and identifying a gene presupposes individuating the gene. I argue that there are at least two levels of the individuation of genes. The transgenic technique can individuate "a gene" as an individual while the technique of gene mapping in classical genetics can only individuate "a gene" as a type or a kind. The two levels of individuation involve different techniques, different objects that are individuated, and different references of the term "gene". Based on the two levels of individuation, I discuss important philosophical implications including the relationship between individuality and individuation and that between individuals and kinds in experimental contexts. I also suggest a new gene conception, calling it "the transgenic conception of the gene."

Keywords: gene concept, individuality, individuation, experiment, classical genetics, transgenic technique

1. Introduction: what is a gene and why individuation matters

"What is a gene?" and its related questions have been asked over and over by philosophers, historians, and scientists of biology (Beurton, Falk, and Rheinberger 2000; Carlson 1991; Falk 1986, 2010; Gerstein et al. 2007; Griffths and Stotz 2006, 2013; Kitcher 1982, 1992; Pearson 2006; Stotz and Griffiths 2004; Snyder and Gerstein 2003; Waters 1994, 2007). Those questions are frequently embedded in discussions about the definition of the term "gene" and the gene concept. As a consequence, the phrase "a gene" in this question usually refers to a type of gene. However, should we use "a gene" to refer to an individual gene, i.e., a gene token? Could it in fact be this?

The question "what is a gene" explicitly implies the problem of identifying genes, and identifying a gene presupposes individuating the gene. In what ways are genes individuated and how do scientists individuate them? I call this *the individuation problem of genes*. This paper shall approach the problem from three different but related perspectives.

From the epistemic perspective, a concept of the gene provides at least a working definition, which by nature is a hypothesis, for scientific research. Any hypothesis of the gene may be in error and may be confirmed only by experimentally individuating particular tokens of some gene. From the semantic perspective, according to a Fregean philosophy of language, the concept of reference usually serves for proper names that refer to individuals or particulars. We may extend the concept of reference to general terms (e. g., "humankind" or "gene kind") for the case in which some token of a kind is presented, and so we use a general term to refer to the kind. This means that at least some token of a kind has to be individuated. This semantic perspective presupposes an ontological perspective: the existence of a kind should be presented or demonstrated by the existence of at least a token of the kind. In the case of the gene, the ontological requirement means that we have to individuate a token of some gene kind. All three perspectives indicate the key status of individuation for answering the question of what a gene is.

According to the literature of analytic metaphysics, "individuation" is understood in a metaphysical and an epistemic sense. In the epistemic sense, someone individuating an object "is to 'single out' that object as a distinct object of perception, thought, or linguistic reference." (Lowe 2005: 75) This epistemic sense presupposes the metaphysical sense, in which what 'individuates' an object "is whatever it is that makes it the single object that it is – whatever it is that makes it one object, distinct from others, and the very object that it is as opposed to any other thing." (Lowe 2005: 75) Bueno, Chen, and Fagan (2018) add a practical sense to the term, interpreting

"individuation" as a practical process through which an individual is produced. They characterize the relation between "individuation" and "individuals" as when "an individual emerges from a process of individuation in the metaphysical sense. Epistemic and practical individuation, then, are processes that aim to uncover stages of that metaphysical process." (Beuno, Chen, and Fagan 2018) The approach to the individuation of genes I adopt herein follows their characterization, especially by focusing on the process of epistemic and practical individuation. Reversely, the case I am investigating in this paper offer an illustration for the new sense of individuation.

Although philosophers have investigated concepts of the gene and its change by examining many cases in scientific practices, they have seldom considered the role that the transgenic technique developed in biotechnology may play in philosophical discussions. This paper explores experimental individuation of genes from the direction of that technique, considering the possibility that a gene is individuated as an individual in the relevant contexts.

This paper thus addresses two central questions: (Q1) In what sense, can we reasonably say that classical geneticists have individuated a gene? (Q2) Are there experiments that can individuate a gene as an individual? Some new questions such as the relationship between individuality and individuation will be derived from the answer to the two questions. This paper is thus structured in the following way.

In the second section, I review the literature about the concepts and references of genes. Section 3 argues that the answer to Q1 is that the geneticists individuate a gene as a type, because they used the chromosomal location technique. Section 4 argues that the answer to Q2 is the experiments that use the transgenic technique. The two answers indicate two different kinds of individuation: individuation of a type and individuation of an individual. This raises a new question about whether or not "individuation of a type" is a consistent phrase. In order to respond to this, section 5 discusses in what sense we individuate a type and compare between two kinds of individuation defined by two different kinds of experiments and techniques: the chromosomal locution of genes and the transgenic experiment. My argument thus involves the relationship between kind and individual in the context of experimentation. Given the new question, Section 6 argues that transgenic experiments can demonstrate a gene type by individuating its tokens, while gene mapping experiments in classical genetics only individuate gene types. Thus, a new gene conception, calling it "the transgenic conception of the gene," can be proposed. I further discuss the relationship among the classical gene concept, the molecular gene concept, and the transgenic conception. In the seventh section, I defend the thesis that practices of individuation in scientific investigations are prior to characteristics of individuality identified by traditionally metaphysical speculations.

2. Concepts and references of the gene

The rapid change of the gene concept has produced a large multitude of gene concepts that have bewildered scientists (Gerstein et. al. 2007; Pearson 2006; Stotz and Griffiths 2004). The confused situation has attracted many philosophers and scientists to provide clarifying analyses. Although scientists as well as philosophers have made endeavors to overcome the predicament, they are motivated differently. Scientists believe that they need a unified concept to help them conduct research and to communicate with each other, because, as developmental geneticist William Gelbert says, "it sometimes [is] very difficult to tell what someone means when they talk about genes because we don't share the same definition" (Pearson 2006: 401). Thus, most scientists seek to redefine the "gene" and tend to adopt a single preferred perspective on the gene concept, although they are well aware with the plurality of gene definitions (Wain et. al. 2002; Gerstein et. al. 2007).

Philosophers at different times have been interested in clarifying concepts of the gene and in investigating the patterns of associated conceptual change. In contrast to actual definitions used by working scientists, they often consider more abstract concepts of the gene that can guide several different definitions in the context of scientific research. Consequently, they conclude that it is almost impossible to find a unified concept of the gene, and hence they take different stances to respond to this situation (rf. Waters 2007). Some are gene skeptics (e.g., Kitcher 1992). Some take a dualistic position, such as Moss (2003), who distinguishes between Gene-P and Gene-D based on the fields in that gene concepts are applied. Some are pluralists, such as Griffiths and Stotz (2006, 2013), who differentiate between three senses of the gene: the instrumental gene, the nominal molecular gene, and the postgenomic molecular gene. Still others are both pluralists and pragmatists. Waters (2018) emphasizes that scientists do and should apply different gene concepts under various investigative contexts.

With some exceptions, few philosophers explore the reference problem of the term "gene". Although Fregean semantics holds that the sense/concept or intension of a name determines its reference or extension, the matter about how a sense determines the reference is not easily seen from the scientific context. The determination of a theoretical term's reference usually involves experimental procedures and techniques that should be investigated and analyzed. Weber (2005, ch.7) does impressive work by providing several reference-determining descriptions of the term "gene" in the history of genetics. Based on those descriptions and the analysis of *Drosophila* genetic practices, he suggests that the pattern of referential change for "gene" is a kind of

freely floating reference. He also argues that different gene concepts refer to *different* natural kinds, which are overlapping but not coextensive.[1] According to Weber, reference for "gene" is fixed in the following manner for classical and molecular genes.

> Reference of [classical] "gene" (2): Whatever (a) is located on a chromosome, (b) segregates according to Mendel's first law, (c) assorts independent of other genes according to Mendel's second law if these other genes are located on a different chromosome, (d) recombines by crossing-over, (e) complements alleles of other genes, and (f) undergoes mutations that cause phenotypic differences. (Weber 2005: 210)

> Reference of [molecular] "gene" (5): The class of DNA sequences that determine the linear sequence of amino acids in a protein. (Weber 2005: 212)

Both classical and molecular gene concepts do refer to natural objects, because, as Weber notes (2005: 210-211), some *tokens* satisfying the reference-determining descriptions are experimentally presented when using the concepts with the intention of referring to sets of entities in historical occasions. However, one should note that the experimented tokens in classical genetics seems to be only some organisms with specific phenotypes (say, fruit flies or other kinds of organisms) while the experimented tokens in molecular biology may be some DNA segments. This difference raises interesting problem: what tokens are individuated in different contexts of experiments?

Before moving to the next section, I want to clarify that the individuation problem of gene concept's tokens is not the issue of gene individuality as raised by Rosenberg (2006: 121-133).[2] He defends the gene individuality thesis in parallel to the species individuality thesis, but Reydon (2009) objects to his argument and defends the gene as a natural kind. This paper aims to discuss how a gene kind and its tokens are individuated rather than whether or not an allele such as *Hbf* (the human fetal hemoglobin gene) is an individual.

3. Chromosomal location of a gene

[1] Baetu (2011: 411) argues that "the referents of classical and molecular gene concepts are coextensive to a higher degree than admitted by Waters and Weber…" However, Baetu builds his argument in terms of Benzer's work on phage. In my view, he does not successfully refute Waters' and Weber's arguments, because the referential change occurred within the classical gene concept, as Weber cogently argues.

[2] Rosenberg uses "natural selection and the individuation of genes" as the title of the section in which he discusses the gene individuality thesis.

Weber's argument indicates that we may and should consider the reference of the classical gene concept independently of the molecular gene concept and others. Weber's reference-determining description of "gene" (2) indicates that the chromosomal location (or mapping) of genes plays a key role in determining referents. However, the question "what tokens are individuated and thus referred to?" does not be answered.

Classical geneticists in the early 20th century located and labeled some specific classical genes on some specific chromosomes. The earliest genetic map (see Figure 1) of *Drosophila melanogaster* (fruit fly) was depicted in 1915. Figure 1 shows that the gene (allele) pair of *Drosophila's* grey body and (mutant) yellow body is located at the first locus on the first chromosome. The second gene pair of red eyes and (mutant) white eyes is located below the grey body gene. The other genes are located below the first two in order. However, every gene is differently distant from the first gene and thus occupies *a single locus* without overlapping. Accordingly, are we able to say that the location of a gene individuates the gene? Before answering this question, it is necessary to discuss how classical geneticists locate a gene on a chromosome. In other words, what technique is used in the process of locating genes?

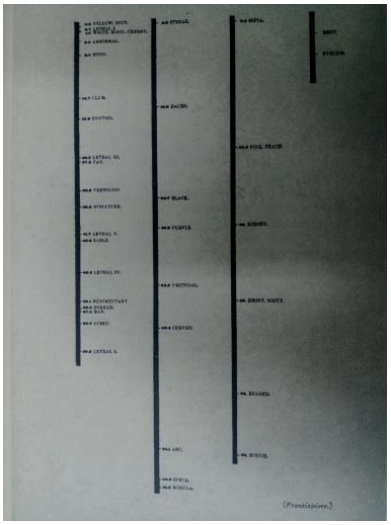


Fig. 1. Genetic map of *Drosophila* in 1915. Reproduced from Morgan, T. H. et. al. (1915).

Chromosomal location or mapping of genes is a well-known story (Darden 1991, Waters 2004, Weber 2005, 2006; Falk 2009). For the purpose of this paper, I introduce a very brief version. In the 1910s, Thomas Hunt Morgan's team developed a

technique to map the linear relations among factors (genes) in linkage groups, using Mendelian breeding data. Morgan and his team discovered that a pair of chromosomes may cross over with each other partially during the period of meiosis. Crossing over produces a specific ratio of the linked traits. Morgan believed that "the percentage of crossing over is an expression of the 'distance' of the factors from each other." (Morgan et.al. 1915: 61) Sturtevant then used percentages of linked characters that exhibited crossing over to calculate the relative positions of the factors to each other. This is the kernel technique for constructing genetic maps. By using genetic maps, Morgan's team determined the loci of many genes on the four chromosomes of *Drosophila*. Given the genetic maps, the classical geneticists assume that no other genes are located at the same position of a chromosome.[3] As a consequence, the single location of a gene actually indicates the individuality of genes.

Genetic maps by nature are diagrammatic models for the actual loci of genes in chromosomes. They are inferences from the statistical data of breeding experiments. Models represent the general. When we say that the location of a gene in a genetic map represents the locus of a classical gene on a chromosome, we really mean that it represents the locus of a type of classical gene on an identical type of chromosome in a cell within a kind of organism. Of course, this implies that a token of a type of classical gene on a token of a type of chromosome can be cognitively identified and discerned, because we can distinguish it from the tokens of the other genes. As a result, we can also count genes within cells. The located genes thus satisfy the two traditional characteristics of individuality: distinguishability and countability.[4]

If all chromosomes were stick-shaped substances of uniform material without complicated structure, then the chromosomal location of classical genes would be able to genuinely individuate them. According to molecular biology, however, chromosomes are a long chain of double helix DNA molecules that curl themselves up in twisted shapes. In such a case, we cannot delineate a located classical gene or depict its contour or boundary, because the chromosomal locus at which the gene is located includes a twisted part of the long DNA molecule. Even by invoking the knowledge from molecular biology, one would still be puzzled by the problem of defining the molecular gene.

4. Individuating molecular genes as individuals

Ever since the era of molecular biology, the continuously accumulating knowledge of genetics has not solved the individuation problem of genes. Instead, it

[3] Of course, a full story is more complicated. For the simplifying purpose, I skip the relevant discussion about gene mutation.

[4] The implications of using these criteria will be discussed in the sixth section.

has brought more troubles about the definition of the gene concept. Is a gene "a sequence of DNA for encoding and producing a polypeptide"? Should we include the start and stop codons (i. e., the regulation problem)? Should we count those introns deleted during the process of transcription into the investigated gene (i.e., the splicing problem)? The difficulty in defining the molecular gene concept directly contributes to the impediment of individuating a gene.

Many gene sequencing projects have been conducted during the genomic era. Scientists do not identify a DNA sequence as a gene and discern the gene from others by using gene sequencing *per se*, because it offers only syntactical orders of genetic codes. Gene annotation, which is used to infer what those annotated sequences do, has been developed to offer *senses* or *intensions* for them. However, the impediment of discerning genes remains, because the definition of the gene is still vague and confusing (rf. Baetu 2012; Gerstein et. al. 2007; Griffiths and Stotz 2013, ch. 4). In fact, gene annotation is based on several assumptions, by which scientists infer that a few sequences may be genes that contribute to phenotypes or functions. Those assumptions need to be confirmed by experimental investigations. Many techniques such as directed deletion, point mutation making, gene silencing, and transgenesis in reverse genetics have been developed to determine what a gene is and what it does (Gilchrist and Haughn 2010).

I argue that the transgenic technique is a very definite and powerful way to individuate a gene. It can even individuate molecular genes as individuals without a clear boundary of a gene or a clear definition of the gene, although the technique is limited.[5] How does the transgenic technique do this? What conditions of individuality allow the technique to individuate a gene as an individual?

Chen (2016) proposes a conception of experimental individuality with three attendant criteria (separability, manipulability, and maintainability of structural unity) and argues that the first experiment of bacteria transformation individuated an antibiotic resistance gene by satisfying the three criteria.[6] Below I reiterate this story in brief.

Stanley Cohen and Herbert Boyer combined DNA of *Escherichia coli* (*E. coli*) in 1973 and 1974 by transferring two different DNA segments encoding proteins for ampicillin and tetracycline resistance into *E. coli*, thereby realizing the transformation of this bacterium (Cohen et. al. 1973; Chang and Cohen 1974). Both DNA segments are called an "antibiotic resistance gene." Cohen and Boyer used small circular

[5] The technique cannot be applied in many occasions because of technological difficulties. It should not be applied to humankind due to ethics consideration. In addition, many gene-modification organisms produced by using the technique may involve ethical issues.

[6] Chen (2016) uses the creation of Bose-Einstein condensates in physical experiments as the other example. Chen's intent is to argue that biological entities and physical entities in laboratories share the same criteria of experimental individuality.

plasmids (extrachromosomal pieces of DNA) as vectors to transfer a foreign DNA segment into a bacterial cell. The plasmids were made by cutting out a (supposed) antibiotic resistance gene from other bacteria with the restriction enzyme *EcoRI*, linking the segment into a plasmid by using another enzyme, DNA ligase. The scientists then transferred the plasmid into an *E. coli* cell without the ability to resist antibiotics. The result, a modified *E. coli* cell, was able to resist antibiotics and contained the antibiotic resistance gene. In that experiment, the antibiotic gene was separated from its original bacteria and then was manipulated (i.e., linked and transferred). Its structural unity was not broken down, hence allowing it to be expressed in the other kind of bacteria. Scientists thus identify it as a gene, an individual biological entity, because the separated, manipulated, and maintained antibiotic gene was naturally separable, manipulable, and maintainable. The photos in Figure 2 show that scientists worked with a single DNA segment, as indicated by (b) in [A] and [B].

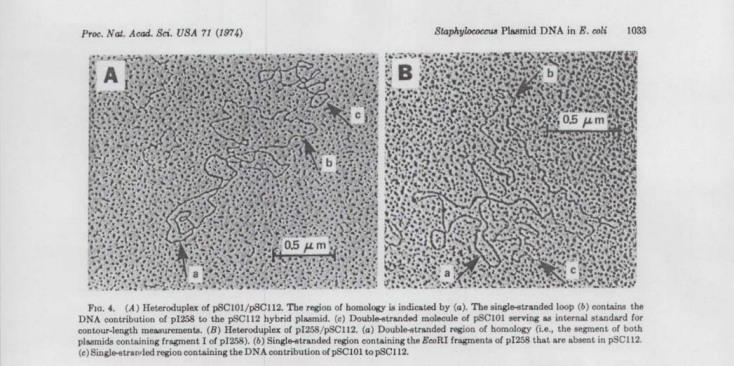


FIG. 4. (*A*) Heteroduplex of pSC101/pSC112. The region of homology is indicated by (*a*). The single-stranded loop (*b*) contains the DNA contribution of pI258 to the pSC112 hybrid plasmid. (*c*) Double-stranded molecule of pSC101 serving as internal standard for contour-length measurements. (*B*) Heteroduplex of pI258/pSC112. (*a*) Double-stranded region of homology (i.e., the segment of both plasmids containing fragment I of pI258). (*b*) Single-stranded region containing the *Eco*RI fragments of pI258 that are absent in pSC112. (*c*) Single-stranded region containing the DNA contribution of pSC101 to pSC112.

Fig. 2. Two pictures of plasmids in bacterial transformation. Reproduced from Chang and Cohen (1974).

I next interpret the performance of the technique used in transgenic experiments as the general process of individuating transgenes. The process has five stages.

(1) Use restriction enzymes to cleave specific segments from recognition sites of long DNA chains. A specific restriction enzyme can cut away a specific DNA segment at a specific site.

(2) Link the cleaved segment of DNA to a plasmid vector by using DNA ligase. The vector is a circular DNA that may come from a wild type of virus.

(3) Incorporate the DNA segment in the vector into the genome of another organism by injecting the plasmid vector to a cell of the target organism. Of course,

they may fail when the intended feature is not expressed.

(4) Make copies of DNA segments by cloning the cell containing the transferred segment of DNA. The aim of DNA cloning is to copy a segment of interest (or a gene) from an organism and produce many copies.

(5) Observe the expression of the novel feature that the target organism does not typically have. If a DNA segment cut from an original organism is successfully pasted into a cell of a target organism and the target organism expresses the intended feature that the original organism has, then one concludes that the segment is a gene.

The first stage corresponds to the separation condition, the second, the third, and the fourth stages to the manipulation condition, and the fifth stage to the maintenance condition. Accordingly, one can easily see that those cut, linked, transferred, pasted, and copied genes are particulars – individuals, because they satisfy the three criteria of experimental individuality that indicates their *singularity* and *particularity*. In other words, a single segment of DNA maintains its structural unity when being separated and manipulated. This is so, because cutting a gene from an original organism is in fact separating it from its environment and because transferring, pasting and copying a gene is manipulating it. If the gene does express the intended feature in a target organism, then this condition indicates that the unity of its chemical and informational structure has been maintained.

5. Two kinds of individuation of genes

The previous discussion indicates that two different objects have been individuated in different experimental and theoretical contexts. In the context of classical genetics, scientists used breeding experiments and theoretical inferences to locate a gene at some locus on a chromosome. They would individuate genes as types if they assume that no other genes could coexist at the same locus. If one interprets the meaning of "individuation" as "only individuals can be individuated," then the phrase "individuating genes as types" sounds unreasonable. Is it better to say "unitization of genes" rather than "individuation of genes"?

It is quite right to say classical geneticists *unitize* genes as types. In a sense, however, we may reasonably say that we individuate a gene as a type, because the type has tokens or members that are distinguishable and countable individuals. Classical geneticists suppose that all types of genes have corpuscular members, i.e., substantive individuals. In such a sense, talking of "individuating genes as types" is reasonable. If no distinguishable and countable members or samples of a kind can be identified, then the kind cannot be individuated. In other words, we cannot individuate

such a kind as water or air that is expressed by "mass" nouns at the macroscopic level, although we can individuate a sample of water by using a container or individuate a water molecule by specific technique at the molecular level. For the cases of experiments using the transgenic technique, molecular biologists physically individuate *singular and particular* gene tokens. Thus, we claim that scientists experimentally individuate genes as individuals in such a context.

In consequence, two different sets of criteria for individuality are presupposed. Experiments using the location technique have individuated a type whose tokens or members are countable individuals rather than matter referred to by mass nouns. In such experimental contexts, we emphasize distinguishability and countability as the indexing features of individuals. Experiments using the transgenic technique individuate singular and particular individuals – gene tokens. For these experimental contexts, we emphasize singularity and particularity of individuals in contrast to universality of types or kinds. We assure the particularity and singularity of the individuals through the realization of experimental individuality, namely, the joint realization of separability, manipulability, and maintainability of structural unity. At this point, more philosophical implications will be discussed in next section.

The two individuated targets indicate two different referential levels of the term "gene" in the literature. As we have seen, when many philosophers and scientists ask "what is a gene," they really refer to a type of gene in conjunction with discussing the gene concept or the definition of "gene." Similarly, in some contexts of scientific investigation, scientists use "a gene" to refer to a type of gene as the phrase "chromosomal location of a gene". In the context of transgenic experiments, however, "a gene" is used to refer to a genuine individual – a single and particular gene token, because scientists have worked with particular objects that maintain their structural unity when being separated and manipulated in the process of experimenting.

The two referential levels indicate two different kinds or levels of experimental individuation, which are realized by two different techniques: the chromosomal location technique and the transgenic technique. Although the two techniques aim to the same target (i.e., genes or types of genes), they physically experiment and manipulate different objects. Experiments using the chromosomal location technique indirectly identify loci of genes by manipulating organisms that contain chromosomes with genes in breeding, while experiments using the transgenic technique directly manipulate DNA segments. Therefore, classical geneticists can only cognitively discern gene types by identifying their loci without practically interacting with gene tokens; they really practically interact with organismal individuals that contain different types of genes. Reversely, molecular biologists can practically interact with gene tokens and then cognitively infer out the existence of a gene type.

6. Gene concepts and individuation

One may still wonder: Can the location technique individuate a singular and particular gene in the sense of individuating entities as individuals? The answer is obviously negative, because that technique cannot separate and manipulate a gene token and maintain its structural unity. On the contrary, one may ask: Can the transgenic technique individuate a type of gene? Here the answer is less clear. In the sense that scientists suppose that a token of a gene has been physically individuated in transgenic experiments, we are allowed to say that the technique also individuates a type of gene. However, scientists are not fully sure that the transgenic technique on a posited gene can be always successfully applied to another individual of the same organism. In fact, the probability of failure is quite high. Unless the experimental individuation of particular tokens can be performed repeatedly and stably, then one can say that the gene tokens indicate a general type of gene and that the type has been identified. However, the object individuated by the technique is not a type of gene, because the technique always requires manipulating particular segments of DNA -- gene tokens. If a kind of transgenic experiment with a specific transgene has been stably repeated, then a type of gene has been discovered by experimentally individuating its tokens in performing such an experiment.

Since transgenic experiments may be successfully and stably performed by using different transgenes, one can extract a special conception of the gene that is characterized by the transgenic technique. I call this "the transgenic conception of the gene," in which *a gene is a transferrable DNA sequence which is able to express a phenotype/function on another kind of organisms*. Of course, this does not imply that those technically untransferrable DNA sequences are not genes, given the fact that the number of transgenes is relatively few to the number of genes located at chromosomes. This is so because scientists do not always find the precise site of a gene (type) and available restriction enzymes to cut the DNA segment of the gene. Thus, the extension of the transgenic conception of the gene is not equivalent to that of the classical gene concept. Due to the limited number of transgenes, the transgenic conception is not yet co-extensional with the molecular gene concept. To be precise, the extension of the former is included within the extension of the latter, because all transgenes are molecular genes but not all molecular genes can be transplanted. In addition, the intension of the transgenic conception is implied in the intension of the molecular gene concept, because the technique was developed from molecular biology. As a consequence, the transgenic conception can be viewed as a *sub-conception* of the molecular gene concept. Nevertheless, we have a conception

derived from scientific practices.

7. The priority of individuation to individuality

Bueno, Chen, and Fagan (2018) promote an approach by which investigating processes of individuation in scientific practices is prior to metaphysical speculation on criteria of individuality. This paper obviously follows the approach. However, this does not mean that we do not need any criterion of individuality in identifying any individual in scientific practices. Rather, criteria of individuality are implied in or extracted from procedures of scientific practices, as the three conditions of experimental individuality are extracted from experimental practices (Chen 2016). Criteria of individuality based on scientific practices may or may not conflict with criteria from metaphysical theories. Considering the relationship between practical criteria and speculative criteria will help us understand practical individuation more deeply.

The metaphysical tradition has identified at least six characteristics or indexing features of individuality in general: particularity, distinguishability, countability, delineability, unity, and persistence (Pradeu 2012: 228-229; Chen 2016: 351).[7] Recently, some philosophers argues that all biological entities are processes (Dupré 2018, Nicholson and Dupré 2018, Pemberton 2018), so I would like to add processuality to the list. Indeed, I believe that all biological individuals pass through a life, i.e., a process (see also Chen 2018), therefore, processuality is a central characteristic of biological individuality. Those characteristics, originally come from metaphysical speculation, can singly, jointly, or collectively serve as epistemic criteria of individuality.

In the context of scientific practices, they are the outcomes from rather than preconditions for the realization of individuation. For example, individuating genes as individuals in the context of transgenic experiments indicates that the separated, manipulated, and maintained genes are particular and singular tokens. As the experimental individuation of gene tokens is realized, those tokens are also distinguishable, countable, unitary, persistent, and passing through a process, because particular and concrete individuals are being separated, manipulated, and maintained. The practices of separation and manipulation indicate epistemic particularity,

[7] Characteristics of individuality can serve as criteria of individuality and thus be involved in a theory of individuation. Bueno, Chen, and Fagan (2018) identify six theories of individuation in traditionally analytic metaphysics. A theory of individuation in the metaphysical sense involves not only "a theoretic construction of the nature of individuality and its attendant criteria," but also other metaphysical concepts such as "property, trope, universal, particular, substance, substratum, time, space, sort or kind." (p. 3) For my purpose, I will discuss only characteristics of individuality rather than any theory of individuation.

distinguishability, and countability. The practice of maintenance of structural unity indicates the unity, persistence, and processuality of the maintained gene token. However, all of the three practices would not indicate the delineation of a gene token, because the spatial boundary of the manipulated gene does not and cannot be delineated. Of course, this point does not mean that delineation is not a characteristic of individuality, but rather that it is not applicable to this case.

Individuating genes as types in classical genetics indicates that the individuated types of genes contain distinguishable and countable tokens, because the individuation is the location of a gene at a chromosome in a diagrammatic model. Supposing that the loci of different genes do not overlap, then the special locus of a gene is thus distinguishable from the locus of another gene. As a consequence, a gene token at a chromosome in a cell of a kind of organism is thus distinguishable from another token of the identical type of gene. All gene types located at chromosomes are countable. Supposing that every organism contains a token of a specific type of gene, then tokens of that gene type are countable. However, chromosomal location of genes does not indicate particular and singular gene tokens, because the individuated objects are only types of genes. As I have argued, the kind of individuation practice did not touch down the manipulation of individuals and remained in the cognitive level which focuses on gene types in general.

Although the concept of individuation can be reasonably applied to a kind whose members are individuals, all characteristics of individuality are not applicable. One cannot apply particularity, delineation, unity, and processuality to gene types, because a gene type is, in principle, universal, occupying multiple spaces, not cohesive, replicable, and non-processual. However, distinguishability and countability can be adequately applied to gene types, because one can distinguish one gene type from another gene type and count gene types when the chromosomal location is realized. In this case, thus, both distinguishability and countability cannot sufficiently demonstrate that the individuated objects are individuals. On the other hand, in the case of transgenic experiments, we can derive particularity, unity, and processuality from the three conditions of experimental individuation (separation, manipulation, and maintenance of structural unity). As a consequence, characteristics of individuality are derived from individuation; they are outcomes of practical individuation.

8. Conclusion

In this paper, I argue that there are at least two kinds of experimental individuation of genes. Scientists individuate genes as types in classical genetics and

individuate genes as tokens in transgenic experiments. Individuating a gene as a type or individuating a gene as an individual depends on the technique used in experimentation. I argue that characteristics of individuality identified in traditional metaphysics are not presupposed by individuation. Rather, they are outcomes or products derived from practical individuation in scientific experiments. I further argue that different kinds of experimental individuation presuppose different concepts of the gene: the classical gene concept and the transgenic conception of the gene. I argue that the transgenic conception can be viewed as a sub-conception of the molecular gene concept. An outstanding problem remains. Whether we can unify different concepts of the gene by integrating different experimental techniques, such as the chromosomal location technique, the technique of genetic sequencing, the techniques in reverse genetics, and the transgenic technique. Future analyses can approach this and other related questions in light of our new understanding of how classical geneticists individuated genes and the role experimental techniques play in identifying a gene as an individual.

Acknowledgment: I thank Alan Love, Ken Water, and Marcel Weber for their very valuable comments and suggestions. This paper has been revised according to their comments.

References

Baetu, Tudor M., 2011. “The referential convergence of gene concepts based on classical and molecular analysis,” *International Studies in the Philosophy of Science*, 24 (4): 411-427.

Baetu, Tudor M., 2012. “Genes after the human genome project.” *Studies in History and Philosophy of Biological and Biomedical Science*, 43: 191-201.

Beurton, P., R. Falk, and H.- J. Rheinberger, 2000. *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*. Cambridge, UK: Cambridge University Press.

Beuno, Otavio, Ruey-Lin Chen, and Melinda B. Fagan, 2018. “Individuation, process, and scientific practice.” In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 1-18. New York: Oxford University Press.

Carlson, E., 1991. “Defining the gene: an evolving concept.” *American Journal of Human Genetics*, 49: 475-487.

Chang, Annie C. Y. and Stanley N. Cohen, 1974. “Genome construction between bacterial species *in vrito*: Replication and expression of *Staphylococcus* plasmids

in *Escherichia coli*," *Proceedings of the National Academy of Science of USA*, 71(4): 1030-1034.

Chen, Ruey-Lin, 2016. "The experimental realization of individuality." In Alexandre Guay and Thomas Pradeu (eds.). *Individuals across the Sciences*, 348-370. New York: Oxford University Press.

Chen, Ruey-Lin, 2018. "Experimental Individuation: Creation and Presentation," In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, . New York: Oxford University Press.

Cohen, Stanley N. et. al., 1973. "Construction of biologically functional bacterial plasmids *in vitro*," *Proceedings of the National Academy of Science of USA*, 70(11): 3240-3244.

Darden, Lindley, 1991. *Theory Chang in Science: Strategies from Mendelian Genetics*. Oxford: Oxford University Press.

Dupré, John, 2018. "Processes, Organisms, Kinds, and Inevitability of Pluralism." In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 25-38. New York: Oxford University Press.

Falk, Raphael, 1986. "What is a gene?" *Studies in History and Philosophy of Science*, 17: 133-173.

Falk, Raphael, 2009. *Genetic Analysis: A History of Genetic Thinking*. Cambridge: Cambridge University Press.

Falk, Raphael, 2010. "What is a gene – revised" *Studies in History and Philosophy of Biological and Biomedical Science*, 41: 396-406.

Gerstein, Mark B. et. al. 2007. "What is a gene, post-ENCODE? History and updated definition." *Genome Research*, 17(6): 669-681.

Gilchrist, Erin and George Haughn, 2010. "Reverse genetics techniques: engineering loss and gain of gene function in plants," *Briefings in Functional Genomes*, 9(2): 103-110.

Griffths, Paul and Karola Stotz, 2006. "Genes in the postgenomic era," *Theoretical Medicine and Bioethics*, 27(6): 253-258.

Griffths, Paul and Karola Stotz, 2013. *Genetics and Philosophy: An Introduction*. Cambridge: Cambridge University Press.

Kitcher, P. S., 1982. "Genes." *British Journal for the Philosophy of Science*, 33: 337-359.

Kitcher, P. S., 1992. "Gene: current usages." In E. Keller and L Lloyd (eds.), *Keywords in Evolutionary Biology*. Cambridge, MA: Harvard University Press, pp. 128-131.

Lowe, E. Jonathan 2005. "Individuation," *The Oxford Handbook of Metaphysics*, ed. Michael J. Loux and Dean W. Zimmerman. Oxford: Oxford University Press.

Maienchein, J., 1992. “Gene: Historical perspectives.” In E. Keller and E. Lloyd (eds.). *Keywords in evolutionary biology*. Cambridge, MA: Harvard University Press, pp. 181-187.

Morgan, Thomas Hunt, et.al., 1915. *The Mechanism of Mendelian Heredity*. New York: Henry Holt and Company.

Moss, Lenny, 2003. *What Genes Can’t Do*. Cambridge, Mass.: The MIT Press.

Nicholson, Daniel J. and John Dupré, 2018. Everything flows: Towards Processual Philosophy of Biology.

Pearson, Helen, 2006. “What is a gene?” *Nature*, 441(25): 399-401.

Pemberton, John. 2018. “Individuating Processes,” In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 39-62. New York: Oxford University Press.

Pradeu, Thomas, 2012. *The Limits of the Self: Immunology and Biological Identity*. Oxford: Oxford University Press.

Reydon, Thomas, 2009. “Gene Names as Proper Names of Individuals: An Assessment.” *British Journal for the Philosophy of Science*, 60(2): 409-432.

Rosenberg, Alexander, 2006. *Darwinian Reductionism*. Chicago: The University of Chicago Press.

Snyder, Michael and Mark Gerstein, 2003. “Defining genes in the genomics era.” *Science*, 300(5617): 258-260.

Stotz, Karola and Paul Griffths, 2004. “Genes: philosophical analyses put to the test.” *History and Philosophy of the Life Sciences*, 26: 5-28.

Wain, H. M., et. al. 2002. “Guidelines for human genome nomenclature,” *Genomics*, 79: 464-470.

Waters, Kenneth C., 1994. “Genes made molecular,” *Philosophy of Science*, 61: 163–185.

Waters, Kenneth C., 2004. “What was classical genetics?” *Studies in History and Philosophy of Science*, 35 (4): 783-809.

Waters, Kenneth C., 2007. “Molecular genetics,” *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/molecular-genetics/

Waters, Kenneth C., 2018. “Don’t Ask ‘What is an individual?’” In Otavio Beuno, Ruey-Lin Chen, and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 91-113. New York: Oxford University Press. (In press)

Weber, Marcel, 2005. *Philosophy of Experimental Biology*. Cambridge, UK: Cambridge University Press.

Weber, Marcel, 2006. “Representing genes: Classical mapping techniques and the growth of genetic knowledge,” *Studies in History and Philosophy of Biological and Biomedical Science*, 29: 295-315.

The Verdict's Out:

Against the Internal View of the Gauge/Gravity Duality

4993 words

Abstract

The gauge/gravity duality and its relation to the possible emergence (in some sense) of gravity from quantum physics has been much discussed. Recently, however, Sebastian De Haro (2017) has argued that the very notion of a duality precludes emergence, given what he calls the internal view of dualities, on which the dual theories are physically equivalent. However, I argue that De Haro's argument for the internal view is not convincing, and we do not have good reasons to adopt it. In turn, I propose we adopt the external view, on which dual theories are not physically equivalent, instead.

1 Introduction

The gauge/gravity duality has generated much discussion about whether space-time geometry or gravity emerges (in some sense) from quantum physics.[1] Recently, however, De Haro [2017] has argued that the very notion of a duality *precludes* the possibility of emergence given what he calls the *internal view* of dualities, on which dual theories are physically equivalent. In turn, this claim impinges upon the broader debate about whether we can make claims about emergence given a duality. After all, since the internal view of dualities is supposed to *rule out* emergence, any such debate is rendered moot once we adopt the internal view. My goal here, though, is to argue that De Haro's argument for the internal view is not convincing. Instead, I propose we adopt the *external view* of dualities, on which dual theories are *not* physically equivalent.

First, I introduce Fraser's [2017] three-pronged distinction of predictive, formal and physical equivalences, characterizing dualities in terms of this distinction (§2.1). I then make things more concrete by briefly considering the gauge/gravity duality via the Ryu-Takayanagi conjecture from the **AdS/CFT** (anti-de Sitter space/conformal field theory) correspondence (§2.2).

Next, I introduce De Haro's interpretive fork between the internal and external views of dualities (§3). I illustrate how the internal view is supposed to preclude emergence, but criticize De Haro's argument for the internal view – that it is meaningless to hold the external view given 'some form of' structural realism and how the two theories are

[1] One prominent physicist who is a proponent of emergent space-time is Seiberg 2007, while philosophers like Rickles 2011/2017, Teh 2013, and Crowther 2014 have all tackled the topic.

'totalizing' in some way – by showing how it does not work without further assumptions (§4). In turn, given the interpretive fork, I propose we adopt the external view instead. In concluding remarks, I briefly discuss this result in relation to the broader debate about emergence within the gauge/gravity duality.

2 Gauge/Gravity through AdS/CFT

2.1 Duality

Fraser [2017] takes two theories related by a duality to have two features: (i) they agree on the transition amplitudes and mass spectra, and (ii) there is a 'translation manual' that allows us to transform a description given by one theory to a description given by another theory. We may explicate (i) and (ii) by first considering distinct sorts of 'equivalence' proposed by Fraser [2017, 35]:

- *Predictive equivalence*: "there is a map from T_1 to T_2 that preserves the values of all expectation values deemed to have empirical significance by T_1 and that preserves the mass spectra, and vice versa."

- *Formal equivalence*: "there is a translation manual from T_1 to T_2 which maps all quantum states and quantum observables deemed to have physical significance by T_1 into quantities in T_2 and respects predictive equivalence, and vice versa."

- *Physical equivalence*: "there is a map from T_1 to T_2 that maps each physically significant quantity in T_1 to a quantity in T_2 with the same physical interpretation and respects both formal and predictive equivalence, and vice versa."

Given our characterization of a duality as (i) and (ii), we may quite naturally say that two theories are dual to one another when they are *predictively* and *formally* equivalent. Furthermore, supposing that this three-pronged distinction exhausts the possible equivalences relevant to physics, we might also say that two theories satisfying (i)-(iii) are also *fully*, or *theoretically*, equivalent.

Here it would be germane to differentiate two distinct sorts of structures in a duality. Given predictive and formal equivalence, the isomorphism holding between physical and empirical quantities of the dual theories suggests a structure, which may be called the *empirical core* of the duality. However, as Teh [2013, 301] also notes, despite the empirical core, "duality is precisely an equivalence between two theories that describe (in general) different physical structures, i.e. theories with non-isomorphic models." In other words, while there is an empirical core, by which physical and empirical quantities are mapped onto one another, these quantities are generally related to other quantities in a quite different manner on each side, viz. there is 'excess structure' exogenous to the empirical core. Without further argument, we are not entitled to 'discard' this 'excess structure', which also means that predictive and formal equivalence (characterizing the empirical core) does not automatically entail physical, and hence full, equivalence.

Given Fraser's framework, I will briefly introduce the gauge/gravity duality more concrete by briefly examining the example of **AdS/CFT** correspondence.

2.2 The AdS/CFT Correspondence

The *gauge/gravity duality*, or *holographic principle*, postulates a duality between a suitably chosen N-dimensional gauge quantum field theory (QFT) that does not describe

gravity, and a quantum theory of gravity in (N+1)-dimensional space-time (the 'bulk') with an N-dimensional 'boundary', on which the gauge theory is defined. Hence the slogan: gauge on the boundary, gravity in the bulk.

The **AdS/CFT** correspondence is a specific case of the gauge/gravity duality. On the one hand, '**AdS**' stands for anti-de Sitter space-time - a maximally symmetric solution to the Einstein equations with a constant negative curvature and a negative cosmological constant. More accurately, though, the '**AdS**' in **AdS/CFT** correspondence should be taken to refer to a string theory of quantum gravity defined *on* a 5-dimensional **AdS**. '**CFT**', on the other hand, refers to a quantum field theory with scale (or conformal) invariance defined on the 4-dimensional boundary of the **AdS**. The **AdS**-side theory is defined in the 'bulk', and the **CFT**-side theory is defined on the 'boundary' of the **AdS** space-time.

The **AdS/CFT** correspondence, then, refers to a postulated duality between the two theories, satisfying (i) and (ii) from §2.1. (i) is satisfied given the postulate that bulk fields propagating in the bulk are coupled to operators in the boundary **CFT**. Hence, the **AdS** theory of gravity will predict exactly the 'same physics', viz. transition amplitudes, expectation values and so on, as the **CFT** theory without gravity.

Beyond empirical, i.e. measurable, quantities, physically significant quantities of **AdS/CFT** must also relate to one another since it is a duality. In other words, (ii) is supposed to hold simply as a core postulate. This is not to say that (ii) is completely unfounded: in particular, we have evidence suggesting that at least *some* physical quantities of dual theories are related to one another in surprising ways, which in turn supports the claim that (ii) holds. Here I will focus on one such relation, the Ryu-Takayanagi conjecture.

The Ryu-Takayanagi conjecture postulates that the entanglement entropy of two regions on the boundary is related to the surface area within the bulk:[2]

$$(\mathbf{RT}):\ S_A\ =\ \frac{Area\left(\widetilde{A}\right)}{4G_N}$$

RT tells us that the entanglement entropy of a region on the boundary of the **AdS**, S_A, viz. the von Neumann entropy[3] in the **CFT**, is directly proportionate (by 4 times the Newtonian gravitational constant) to the area of the boundary surface $\widetilde{A}$ bisecting the bulk, dividing the two entangled regions on the boundary. Below, *Fig. 1.* shows a simplified diagram for visualizing **RT**.

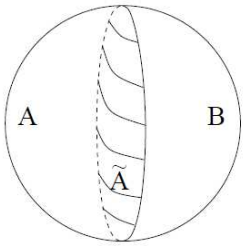


Fig. 1. The area $\widetilde{A}$ bisects the bulk space-time into two, and on the boundaries of the two parts we define the regions A and B. The Ryu-Takayanagi formula tells us that given a change in S_A we get a change in the size of $\widetilde{A}$ by the proportion of $\frac{1}{4G_N}$. [Figure taken from Van Raamsdonk 2010]

RT paints an interesting picture for emergence of space-time geometry from quantum theory: the area of a space-time itself is closely related to quantum entanglement entropy in a surprising way. An increase in the entanglement entropy between two

[2]See Ryu & Takayanagi 2006 for technical details.

[3]The von Neumann entropy is given by $S_A = -Tr(\rho_A log \rho_A)$. The reduced density matrix describing the region A, ρ_A, is obtained from tracing over the B-components of the combined density matrix of A and the entangled region B, ρ_{AB}: $\rho_A = Tr_B(\rho_{AB})$.

regions of a field described by **CFT** leads to a proportionately increasing boundary area of the bulk, and hence a geometric (or gravitational) phenomenon is described in terms of a quantum phenomenon.[4]

Given relations like **RT**, we can also see more clearly how **AdS/CFT** is supposed to satisfy (ii): physically significant quantities, such as 'area' of space-time in the bulk and 'entanglement entropy' between two regions on the boundary, are mapped to one another via suitable equations. Hence, **AdS/CFT** is a special case of the gauge/gravity duality: a theory of quantum gravity on a $(N+1)$-dimensional **AdS** space-time is dual to a **CFT** defined on its N-dimensional boundary.

With the gauge/gravity duality made concrete, let us turn to the interpretive task.

3 The Internal View

Dieks et al. [2015] and De Haro [2017] proposes an interpretive fork for dualities: we can either adopt an internal or external view. De Haro describes the *internal view* as such:

> if the meaning of the symbols is not fixed beforehand, then the two theories, related by the duality, can describe the same physical quantities. [...] we have two formulations of one theory, not two theories. [De Haro 2017, 116]

On the contrary, the *external view* holds that:

> the interpretative apparatus for the entire theory is fixed on each side. [...] On this interpretation there is only a formal/theoretical, but no empirical, equivalence between the two theories, as they clearly use different physical

[4]See Van Raamsdonk 2010 for an excellent summary of this picture.

> quantities; only one of them can adequately describe the relevant empirical observations.

Is De Haro's characterization of the external view adequate? The fact that there is no 'empirical' equivalence (what Fraser calls physical equivalence) between two theories does not entail that at most one of them can adequately describe the relevant empirical observations, where one description is 'correct' and the other 'wrong', nor does it entail mutually exclusive physics where only one theory can be correct at any one time. To assume so seems to rule out, by fiat, the possibility of emergence, since emergence relies on *both* theories being in a way adequately descriptive of the world (except one is more 'fundamental' than the other). Hence, taking in account Fraser's framework, I re-characterize the external view as such: it is simply the claim that the two dual theories are *physically non-equivalent* i.e. have distinct physical interpretations, despite formal and predictive equivalence.

Given the interpretive fork, if we are led to forsake the internal view, then we are motivated to accept the external view instead. As such, my strategy here is to show that we should forsake the internal view, and in turn accept the external view instead.

To better understand what the internal view is claiming, I break it down into three constituent claims.

The first claim is that of *theoretical equivalence*: under the gauge/gravity duality, the two theories (e.g. **AdS** and **CFT**) are taken to be simply different formulations of a single theory, describing the same physical quantities despite their obvious differences. As Dieks et al. puts it, 'the two theories collapse into one' [2015, 209-210]. In light of Fraser's framework described in §2.3, this claim means that the gauge/gravity duality, on

the internal view, involves the conjunction of predictive, formal and physical equivalences. In other words, beyond a one-to-one mapping (a 'translation manual') of relevant physical quantities and the sharing of all transition amplitudes, mass spectra and other observable predictions, the internal view claims that the two theories also have the *same physical interpretation*. However, as Fraser [2017, 35] notes, "predictive equivalence does not entail formal equivalence, and formal equivalence does not entail physical equivalence." Formal and predictive equivalence cannot entail physical equivalence on their own.

The internal view's claim of theoretical equivalence, then, must require an additional claim of *physical equivalence*, in addition to formal and predictive equivalence: the dual theories are taken to be physically equivalent, and hence have the same physical interpretation. As per §2.1, this would indeed entail theoretical equivalence.

Physical equivalence is in turn justified by a third claim, that the two theories in a duality should be left *uninterpreted*. As De Haro claims above, assume 'the meaning of the symbols is not fixed beforehand'. Then, given formal and predictive equivalence, we have an isomorphism between the dual theories' (now-uninterpreted) 'physical quantities' and numerical predictions, viz. an uninterpreted empirical core. Ignoring the 'excess structure' exogenous to the empirical core, we can then take the empirical core to be representing a single uninterpreted theory, where the now-uninterpreted 'quantities' of each dual theory now refer to the 'places' or 'nodes' of the empirical core's structure. As Dieks et al. (2015) puts it,

> A in one theory will denote exactly the same physical quantity as B [...] if these quantities occupy structurally identical nodes in their respective webs

> of observables and assume the same (expectation) values. [Dieks et al. 2015, 209]

Now, given this situation, it might seem plausible to claim that the dual theories are really physically equivalent. Consider **RT**. On the internal view, we are led to say that 'area' really has the same meaning as 'entanglement entropy'. After all, in the theoretical structure that is supposed to matter on the internal view, viz. the empirical core, the two terms are related structurally in the same way to other terms elsewhere (sans a proportional constant). Given that the two theories is also stripped of all prior physical meaning, this structural identity suggests that the 'area' and 'entanglement entropy' are really describing the same quantities, despite their obvious non-isomorphism more generally (e.g. different equations in computing these quantities in their respective theories, the terms involved in calculating them, and so on). In other words, it seems that we are allowed to proclaim physical equivalence on this view.

If we do accept this third claim, we get physical and hence theoretical equivalence, and so the internal view does preclude the possibility of emergence: Theoretical equivalence effectively rules out any account of emergence. If the two dual theories are really just different formulations of one theory, then there is nothing for this new, unified, theory to emerge *from*: nothing can emerge from itself in any interesting way. Subsequently, a duality is supposed to *preclude* emergence on the internal view.

Agreed: physical equivalence entails theoretical equivalence, and theoretical equivalence rules out any sort of emergence. However, are we forced to adopt physical equivalence given the internal view? De Haro himself seems unclear on this point. Note the use of "can" in his characterization of the internal view above: "the two theories,

related by the duality, *can* describe the same physical quantities" [2017, 116, emphasis mine]. Are we supposed to believe that physical equivalence *can* hold, or that it *must* hold, on the internal view? In other words, since physical equivalence hangs on the third claim of leaving terms of the dual theories uninterpreted, *must* we adopt the third claim, or is it merely *possible*?

De Haro seems to suggest that theoretical, and hence physical, equivalence *must* hold, since he assumes the two dual theories to be 'two formulations of *one theory*' [emphasis mine]. However, later on, he suggests that physical equivalence merely *can* hold, when he considers an example of leaving dual theories uninterpreted beyond structural relations:

> For what might intuitively be interpreted as a 'length, a reinterpretation in terms of 'renormalisation group scale is now *available*.[5] [De Haro 2017, 116, emphasis mine]

The *availability* of an interpretative stance – in our case of **RT**, of interpreting bulk boundary surface area to be the same physical quantity as entanglement entropy – surely does not entail the *necessity* of the stance. Hence, there are two readings of the internal view: on the weak reading, we take the modal talk – e.g. a reinterpretation being 'available' or how we 'can' describe the same physical quantities – seriously, and on the strong reading we ignore the modal talk completely.

On the one hand, the claim that the internal view precludes emergence is not true on the weaker view. On this view, *if* we assume that the terms on both sides of the duality are uninterpreted, then there is no emergence; *but* this is not forced on us. In turn, this

[5]For context, though unmentioned in this paper, length and renormalisation group scale are also dual quantities in AdS/CFT.

makes the preclusion of emergence merely possible. However, this reading of the internal view does not rule out emergence as De Haro claims. I will thus assume that De Haro intends for us to take the strong reading of the internal view, which does claim that the terms of the both sides *are* uninterpreted.

However, we have not yet seen a compelling reason for accepting the claim that we *have to* see the terms of the dual theories as uninterpreted, and subsequently that physical equivalence *must* hold. *A fortiori* we are not obliged to accept the internal view.

Indeed, something is odd about the argument structure I mapped out: To establish the second claim of physical equivalence, we must establish the third claim, that we must discard anything beyond the empirical core and to leave the terms uninterpreted. However, to justify leaving the terms uninterpreted requires a convincing argument for assuming physical equivalence between the two theories to begin with! Otherwise, we have no reason to simply discard the 'excess' structure and leave the dual theories' terms uninterpreted.

Hence, further arguments are required to establish the third claim. Furthermore, if we discover that this argument is wanting, we shall then have reasons to reject the internal view.

4 De Haro's Argument

De Haro does provide an argument, which runs on the idea that two plausible commitments entails the internal view: the commitment that the dual theories are theories of the whole world in some suitably totalizing manner, and the commitment to "some form of structural realism" [2017, 116].

Let us begin by examining the two commitments. The first commitment implies that dual theories are theories of the whole world, in the sense that they are "both candidate descriptions of the same world" [Dieks et al. 2015, 14]. However, *prima facie* this is not true, since on one hand we have a theory of gravity/space-time geometry, while on the other we have a theory without (not to mention different dimensionalities). How can two theories, one describing something the other does not, both be about the same world? We can try to make this assumption intelligible by taking into account the translation manual between the two theories. Given the translation manual, we can claim that the **CFT** theory without gravity does describe gravity in a way. Consider **RT**: while the entanglement entropy described within **CFT** does not appear to describe space-time geometry *by itself*, the **CFT** plus the translation manual *and* **AdS** (in this case **RT**) *does* describe space-time geometry, albeit in a higher-dimensional space-time. When the entanglement described within the **CFT** changes, the boundary surface area in the **AdS**-side theory with gravity changes as well. Hence, by considering the translation manual given by the duality, the first commitment is made plausible.

The second commitment requires us to adopt some form of structural realism. Structural realism here can be understood loosely, since nothing turns on the particular account of structural realism we employ. Furthermore, De Haro himself does not specify precisely what he means by 'some form of' structural realism. As such, I will likewise adopt a loose notion of structural realism: I understand it to be the view that we should be (metaphysically or epistemically) committed only to the mathematical or formal structure of our theories, and this entails, among other things, that theoretical terms are to be defined in terms of their relations to other places or nodes in this formal structure.

Now, De Haro then claims that the two commitments entail the internal view:

> If [the two commitments] are met, it is impossible, in fact meaningless, to decide that one formulation of the theory is superior, since both theories are equally successful by all epistemic criteria one should apply. [De Haro 2017, 116]

Since he does not flesh out his argument in much detail, I attempt to reconstruct his argument in a plausible fashion: firstly, let us grant the two commitments. Do these commitments commit us to the conclusion that it is meaningless to differentiate between the two dual theories?

Dieks et al. [2015, 209] claims that given the first commitment, "it is no longer clear that there exists an 'external' point of view that independently fixes the meanings of terms in the two theories". However, I must admit I do not see why this is the case: as I explained above, the first commitment only makes sense *if* we understand both theories as having pre-determined meanings, and *then* relating them via the duality/translation manual. In other words, the first commitment is perfectly compatible with the external view.

For the remainder of this paper I focus on the second commitment instead. I think the second commitment *does* entail that differentiating the two theories is meaningless, *only if* we believe that one should be a structural realist (epistemically/metaphysically) only about the empirical core of the duality, discarding the 'excess structure' which made the two theories distinct structures to begin with. In other words, we want to say that this 'excess structure' was not physically significant to begin with: only the empirical core was relevant to physics. It seems that this is required to make sense of the claim that it is 'meaningless' to say that one formulation, e.g. the **CFT** side, is better than the

other, e.g. the **AdS** side. If structural realism commits us only to the empirical core of the dual theories, then accordingly there is really only one structure in question. Hence, it is meaningless to ask which structure is better (there is only one). If there is only one structure, then the internal view seems to hold: under a structural realist view, the terms of the dual theories are defined in terms of their places in the structure. Hence, within the empirical core's structure, the different terms of the dual theories really mean the same thing, and hence we get some version of the internal view.

Why should we, even as structural realists, commit ourselves only to the empirical core? The argument seems to me to be an epistemic one: we should believe that the structure relevant to the two theories given the duality must really be common to both theories because, as De Haro claims above, "both theories are equally successful" by all epistemic criteria we apply. If this is true then it seems we have no way of differentiating between the two theories, and the best explanation for this epistemic equivalence is to appeal to their being 'the same' in some way. The only thing in common between the dual theories is the empirical core, so we should take this to be what explains their epistemic equivalence. Everything else (i.e. the 'excess structure') can be discarded, since they are irrelevant differences. As such, structural realism should commit us only to the empirical core.

However, it is not clear that the dual theories are indeed epistemically equivalent. In a naive sense, they are epistemically equivalent if one takes 'epistemic' to be 'empirical' equivalence. Given the duality, i.e. formal and predictive equivalence, it is trivial that the two theories are also 'empirically' equivalent. However, I do not think such a notion of empirical equivalence *exhausts* the epistemic criteria for differentiating between scientific theories. Of course, one main desideratum for scientific theorizing is to provide

predictions, descriptions and explanations of phenomena. Beyond that, though, I contend that another desideratum of scientific theorizing is to look for ways to develop better scientific theories, be it a more unificatory theory, a more explanatory theory, and so on.

We see this in play when De Haro discusses the position/momentum duality in quantum mechanics: "this duality is usually seen as teaching us something new about the nature of reality: namely, that atoms are neither particles, nor waves. By analogy, it is to be expected that gauge/gravity dualities teach us something about the nature of spacetime and gravity" [2017, 117]. However, this is only possible *if* the two theories were not epistemically equivalent! If they were epistemically equivalent, then how could we learn anything new from one theory that we cannot already learn from another? If 'area' and 'entanglement entropy' really meant the same thing and had the same physical interpretation, how could we learn something new when we realize that area can be related (via **RT**) to quantum entanglement? Indeed, this criticism extends generally to the internal view: how can we learn anything new from a duality if the dual theories are just the 'same theory', and indeed are *uninterpreted* to begin with? We learn something new when two *different* things are related in a surprising way, *especially* when they are related to other quantities, on each side, in interesting ways; I do not see how we can learn something new when one and the same thing is related to itself.

Furthermore, the two theories are *not* epistemically equivalent when we consider the methodological concerns of physicists, who generally note that the **CFT** is well-understood, while the dual string theory of gravity is not. For example, Horowitz and Polchinski [2009] notes that we only approximately understand the gravitational theory, but the **CFT** has been developed to very precise degrees. Lin points out that:

> A dictionary is reasonably well developed in the direction of using classical gravity to study the **CFT**, but the converse problem how to organize the information in certain **CFT**'s into a theory of quantum gravity with a semi-classical limit is hardly understood at all. [2015, 11]

If both theories are equally successful by *all* epistemic criteria we have, then this situation should not appear. Rather, it seems that scientific practice is of the opinion that the two theories are, in fact, *not* epistemically equal: one is more successful than the other in terms of a variety of criteria, such as precision of calculation, ease of understanding, availability of a non-perturbative analysis, and so on. It is one reason why **AdS/CFT** is such an interesting area of research: it allows us to understand a hard-to-understand theory in terms of an easier-to-understand theory. Unless one is given arguments for why such criteria should *not* be epistemically relevant, the dual theories, I contend, are *not* epistemically equivalent.

Of course, one could assume that the *goal* or *ideal*, when we fully understand the translation manual, is to render both theories equally epistemically successful. However, this presumes that both sides *will* end up being just as easy to compute, or understand, and so on. Of course, if we do discover a more fundamental characterization of *why* the two dual theories are related by the duality as such, e.g. the sort of 'deeper' theory Rickles [2011, 2017] hopes for, then clearly we are entitled to the internal view since this 'deeper' theory will ideally explain why the dual theories, despite their apparent differences, can be seen as different facets of a single theory, just like how special relativity unified electromagnetism and made it plausible to understand both the electric and magnetic fields as facets of the 'deeper' Faraday tensor field. Right now, though,

there is no such theory in sight, making this point inadequate for supporting the internal view.

Given the foregoing, it is not clear there is epistemic equivalence: the epistemic argument does not hold. The upshot is that we are not compelled to provide an explanation for why the dual theories are epistemically equivalent to begin with (they are not), and hence we have no need to commit ourselves only to the common empirical core, *even* as structural realists, nor to think that differentiating the dual theories is meaningless.

Recall the oddity I pointed out in §3, though. The claim of physical equivalence hangs on leaving the dual theories uninterpreted, but this latter claim was itself motivated by physical equivalence. It was hoped, then, that the epistemic argument could provide independent motivation for adopting physical equivalence. Given my criticism of De Haro's additional argument, though, the circle returns, and leaves the two claims uncompelling. Hence, we should not adopt the internal view itself. Furthermore, my criticisms suggest that the dual theories are in fact *not* epistemically equivalent, and this suggests that the default stance is one where the two theories are not theoretically equivalent at all. Given the duality, the only way this can be so is to adopt the view that the dual theories are physically non-equivalent; in other words we should adopt the external view instead.

To conclude, given the dialectic set up by the interpretive fork, and the inadequacies of the internal view, I suggest that we adopt the external view instead.

5 The Way Forward

Let me end by commenting on the external view and the broader debate on whether there is emergence given a duality (§1). In §3 we have seen how the internal view precludes emergence simply because there are no two distinct theories to speak of: we merely have two ways of looking at a single theory. This in turn swiftly rules out any talk of emergence. The external view, though, does not rule out emergence quite so easily, and there is some leeway to speak of emergence since we *do* have two distinct theories which are, as Teh noted, generically *not* isomorphic to one another. However, given the formal and predictive equivalences demanded by a duality relation, a duality relation is symmetric, and so there is nothing within a duality that will formally broker the asymmetry between two theories we often associate with emergence. One way to do so, as Teh (2013) suggests, is to introduce a claim of relative fundamentality, i.e. which theory is 'more fundamental' than another, is required to break the symmetry and provide us with the required asymmetry for emergence. While the external view does not entail this, it does not rule it out either. Hence, the external view does not preclude emergence; instead, it directs attention about emergence and duality away from the interpretative fork, onto whether and how one can make claims about relative fundamentality in the context of dualities. Alas, this requires much more attention than I can afford here: I leave it for another day.

References

Dieks, D., J. van Dongen, and S. D. Haro (2015). Emergence in Holographic Scenarios for Gravity. *Studies in the History and Philosophy of Modern Physics 52*, 203–216. 10.1016/j.shpsb.2015.07.007.

Fraser, D. (2017). Formal and Physical Equivalence in Two Cases in Contemporary Quantum Physics. *Studies in the History and Philosophy of Modern Physics 59*, 30–43. 10.1016/j.shpsb.2015.07.005.

Haro, S. D. (2017). Dualities and Emergent Gravity: Gauge/Gravity Duality. *Studies in the History and Philosophy of Modern Physics 59*, 109–125. DOI: 10.1016/j.shpsb.2015.08.004.

Haro, S. D., N. Teh, and J. Butterfield (2017). Comparing Dualities and Gauge Symmetries. *Studies in the History and Philosophy of Modern Physics 59*, 68–80. DOI: 10.1016/j.shpsb.2016.03.001.

Horowitz, G. and J. Polchinski (2009). Gauge/gravity duality. In D. Oriti (Ed.), *Approaches to quantum gravity: Toward a new understanding of space time and matter*, pp. 169–186. Cambridge: Cambridge University Press. arXiv:gr-qc/0602037.

Raamsdonk, M. V. (2010). Building up spacetime with quantum entanglement. *General Relativity and Gravitation 42*(10), 2323–2329. 10.1007/s10714-010-1034-0.

Rickles, D. (2011). A Philosopher Looks at Dualities. *Studies in the History and Philosophy of Modern Physics 42*(1), 54–67. DOI: 10.1016/j.shpsb.2010.12.005.

Rickles, D. (2017). Dual Theories: 'Same but Different or 'Different but Same? *Studies in the History and Philosophy of Modern Physics 59*, 62–67. 10.1016/j.shpsb.2015.09.005.

Ryu, S. and T. Takayanagi (2006). Holographic Derivation of Entanglement Entropy from AdS/CFT. *Phys. Rev. Lett 96*(18). 10.1103/PhysRevLett.96.181602.

Seiberg, N. (2007). Emergent spacetime. In D. Gross, M. Henneaux, and A. Sevrin (Eds.), *The Quantum Structure of Space and Time*, pp. 163–178. Singapore: World Scientific. DOI: 10.1142/9789812706768_0005.

Teh, N. (2013). Holography and Emergence. *Studies in the History and Philosophy of Modern Physics 44*(3), 300–311. DOI: 10.1016/j.shpsb.2013.02.006.

Causal Discovery and the Problem of Psychological Interventions

PSA 2018, Seattle

Markus Eronen

University of Groningen

m.i.eronen@rug.nl

Abstract

Finding causes is a central goal in psychological research. In this paper, I argue that the search for psychological causes faces great obstacles, drawing from the interventionist theory of causation. First, psychological interventions are likely to be both fat-handed and soft, and there are currently no conceptual tools for making causal inferences based on such interventions. Second, holding possible confounders fixed seems to be realistically possible only at the group level, but group-level findings do not allow inferences to individual-level causal relationships. I also consider the implications of these problems, as well as possible ways forward for psychological research.

1. Introduction

A key objective in psychological research is to distinguish causal relationships from mere correlations (Kendler and Campbell 2009; Pearl 2009; Shadish and Sullivan 2012). For example, psychologists want to know whether having negative thoughts is a cause of anxiety instead of just being correlated with it: If the relationship is causal, then the two are not just spuriously hanging together, and intervening on negative thinking is actually one way of reducing anxiety in patients suffering from anxiety disorders. However, to what extent is it actually possible to find psychological causes? In this paper, I will seek an answer this question from the perspective of state-of-the-art philosophy of science.

In philosophy of science, the standard approach to causal discovery is currently interventionism, which is a very general and powerful framework that provides an account of the features of causal relationships, what distinguishes them from mere correlations, and what kind of knowledge is needed to infer them (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009, Woodward 2003, 2015b; Woodward & Hitchcock 2003). Interventionism has its roots in Directed Acyclic Graphs (DAGs), also known as causal Bayes nets, which are graphical representations of causal relationships based on conditional independence relations (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009). More recently, James Woodward has developed interventionism into a full-blown philosophical account of causation, which has become popular in philosophy and the sciences. Several authors have also argued that interventionism adequately captures the role of causal thinking and reasoning in psychological research (Campbell 2007; Kendler and Campbell 2009; Rescorla forthcoming; Woodward 2008).

Based on interventionism, I will argue in this paper that the discovery of psychological causes faces great obstacles. This is due to problems in performing psychological interventions and deriving interventionist causal knowledge from psychological data.[1] Importantly, my focus is not on the existence or possibility of psychological causation, but on the *discovery* of psychological causes, which is a topic that has so far received little attention in philosophy.[2] Although I rely on interventionism, my arguments are based on rather general principles of causal inference and reasoning in science, and will thus apply to any other theory of causation that does justice to such principles.

The focus in this paper will be on the discovery *individual-level* (or within-subject) causes, not *population-level* (or between-subjects) causes. The first refers to causal relationships that hold for a particular individual: for example, John's negative thoughts cause John's problems of concentration. The latter refers to causal relationships that obtain in the population as a whole: for example, negative thoughts cause problems of concentration in a population of university students. It widely thought that ultimate goal of causal inference is to find individual-level causes, and that a population-level causal relationship should be seen as just an average of individual-level causal relationships (Holland 1986): For example, the causal relationship between negative thoughts and problems of concentration in a population of university students is only interesting insofar as it *also* applies to at least some of the individual students in the

[1] See Eberhardt (2013; 2014) for different (and domain-independent) problems for interventionist causal discovery.
[2] There is an extensive debate on the question whether interventionism vindicates non-reductive psychological causation by providing a solution the causal exclusion problem (e.g., Baumgartner 2009, Eronen 2012, Raatikainen 2010, Woodward 2015). I will sidestep this debate here, as my focus is not on the existence of non-reductive psychological causation, but on the discovery of psychological causes, be they reducible or not.

population.[3] Thus, in this paper I will discuss population-level causal relationships only when they are relevant to discovering individual-level causes.

Importantly, the distinction between population-level and individual-level causation is different from the distinction between type and token causation, even though the two distinctions are sometimes mixed up in the philosophical literature (see also Illari & Russo 2014, ch. 5). Token causation refers to causation between two actual events, whereas type causation refers to causal relationships that hold more generally. Individual-level causes can be either type causes or token causes. An example of an individual and type causal relationship would be that John's pessimistic thoughts cause John's problems of concentration: This is a general relationship between two variables, and not a relationship between two actual events. An example of an individual and token causal relationship would be that John's pessimistic thoughts before the exam on Friday at 2 pm caused his problems of concentration in the exam. As interventionism is a type-level theory of causation, and the aim of psychological research is primarily to discover regularities, not explanations to particular events, in this paper I will only discuss the discovery of type (individual) causes.

The structure of this paper is as follows. I will start by giving a brief introduction to interventionism, and then turn to problems of interventionist causal inference in psychology: First, to problems related to psychological interventions (section 2), and then to problems arising from the requirement to "hold fixed" possible confounders (section 3). After this, I will consider the possibility of the inferring psychological causes without interventions (section 4). In the last

[3] It has been argued that population-level (between-persons) causal relationships can also be real without applying to any individual (Borsboom, Mellenbergh, and van Heerden 2003). However, also those who believe in these kind of population-level causes agree that discovering individual causes is an important goal as well.

section, I discuss ways forward and various implications that my arguments have for psychology and its philosophy.

2. Interventionism

Interventionism is a theory of causation that aims at elucidating the role of causal thinking in science, and defining a notion of causation that captures the difference between causal relationships and mere correlations (Woodward 2003). Thus, the goal of interventionism is to provide a methodologically fruitful account of causation, and *not* to reduce causation to non-causal notions or analyse the metaphysical nature of causation (Woodward 2015b). In a nutshell, interventionist causation is defined as follows:

X is a cause of Y (in variable set **V**) if and only if it is possible to *intervene* on X to change Y when all other variables (in **V**) that are not on the path from X to Y are *held fixed* to some value (Woodward 2003).

Thus, in order to establish that X is a cause of Y, we need evidence that there is some way of intervening on X that results in a change in Y, when off-path variables are held fixed.[4] Importantly, it is not necessary to actually perform an intervention: What is necessary is knowledge on what *would* happen if we *were* to make the right kind of intervention.

[4] More precisely, this is the definition for a *contributing* cause. X is a *direct* cause of Y if and only if it is possible to intervene on X to change Y when all other variables (in **V**) are held fixed to some value (Woodward 2003). Thus, the definition of a contributing cause allows there to be other variables on the causal path between X and Y, whereas the definition of a direct cause does not. This does not reflect any substantive metaphysical distinction, as the question whether X is a direct or contributing cause is relative to what variables are included in the variable set. Importantly, notion of a contributing cause is *not* relative to a variable set in any strong sense – if X is a cause of Y in some variable set, then X will be a cause of Y in all variable sets where X and Y appear (Woodward 2008b). This is because the definition of an intervention is not relativized to a variable set.

The notion of an intervention plays a fundamental role in the account, and is very specifically defined. Here is a concise description of the four conditions that an intervention has to satisfy (Woodward 2003).

Variable *I* is an intervention variable for X with respect to Y if and only if:

(I1) *I causes* the change in X;

(I2) The change in X is *entirely* due to *I* and not any other factors;

(I3) *I* is not a cause of Y, or any cause of Y that is not on the path from X to Y;

(I4) *I* is *uncorrelated* with any causes of Y that are not on the path from X to Y.

The rationale behind these conditions is that if the intervention does not satisfy them, then one is not warranted to conclude that the change in Y was (only) due to the intervention on X. Thus, in simpler terms, the intervention should be such that it changes the value of the target variable *X* in such a way that the change in *Y* is *only* due to the change in *X* and not any other influences (Woodward 2015b). For example, if the intervention is correlated with some other cause of Y, say Z, that is not on the path from X to Y (violating I4), then the change in Y may have been (partly) due to Z, and not just due to X. Following standard terminology in the literature, I will call interventions that satisfy the criteria I1-I4 *ideal* interventions. I will now go through various problems in performing ideal interventions in psychology, starting from problems related to conditions I2 and I3 (section 3), and then turn to problems related to I4 and the "holding fixed" part of the definition of causation (section 4).

3. Psychological interventions

Before discussing psychological interventions, an important distinction needs to be made: The distinction between relationships where (1) the cause is *non-psychological*, and the *effect* is psychological, and (2) where the *cause* (and possibly also the effect) is *psychological*.[5] A large proportion (perhaps the majority) of experiments in psychology involve relationships of the first kind: The intervention targets a non-psychological variable (X) such as medication vs. placebo, therapy regime vs. no therapy, or distressing vs. neutral video, and the psychological effect of the manipulation of this non-psychological variable is tracked. In other words, the putative causal relation is between a non-psychological cause variable (X) and a psychological effect variable (Y). In these cases, it is possible to do (nearly) ideal interventions on the putative cause variable (X) by ensuring that the change in X was caused (only) by the intervention, that the intervention did not change Y directly, and that it was uncorrelated with other causes of Y. It is of course far from trivial to make sure that these conditions were satisfied, but as the variables intervened upon are non-psychological, making the right kinds of interventions is in principle not more difficult than in other fields. As regards the psychological effect variable (Y), there is no need to intervene on it; it is enough to measure the change in Y (which, again, is far from trivial, but faces just the usual problems in psychological measurement, which will be discussed below). The fact that many psychological experiments involve this kind of causal relationships may have contributed to the recent optimism on the prospects of interventionist causal inference in psychology.

[5] The line between psychological and non-psychological variables is likely to be blurry. However, for the present purposes it is not crucial where exactly the line should be drawn: My arguments apply to cases where it is clear that the cause variable is psychological (such as the examples in the main text), and such cases abound in psychological research.

However, psychological research also often concerns relationships of the second kind, that is, relationships where the *cause* is psychological. This is, for example, the case when the aim is to uncover psychological mechanisms that explain cognition and behavior (e.g., Bechtel 2008, Piccinini & Craver 2011), or to find networks of causally interacting emotions or symptoms (e.g., Borsboom & Cramer 2013). The reason why these relationships are crucially different from relationships of the first kind is that now the variable intervened upon is psychological, so the conditions on interventions now have to be applied to psychological variables.

Ideal interventions on psychological variables are rarely if ever possible. One reason for this has been extensively discussed by John Campbell (2007): Psychological interventions seem to be "soft", meaning that the value of the target variable *X* is not completely determined by the intervention (Eberhardt & Scheines 2007; see also Kendler and Campbell 2009; Korb and Nyberg 2006). In other words, the intervention does not "cut off" all causal arrows ending at *X*. As a non-psychological example, when studying shopping behaviour during one month by intervening on income, an ideal intervention would fully determine the exact income that subjects have that month, whereas simply giving the subjects an *extra* 5000€ would count as a soft intervention (Eberhardt & Scheines 2007). Similarly, if we intervene on John's psychological variable *alertness* by shouting "WATCH OUT!", this does not completely cut off the causal contribution of other psychological variables that may influence John's *alertness*, but merely adds something on top of those causal contributions (Campbell 2007). As most or all interventions on psychological variables are likely to be soft, Campbell proposes that we should simply allow such soft interventions in the context of psychology. Campbell argues that these kind of interventions can still be informative and indicative of causal relationships (Campbell

2007), and this conclusion is supported by independent work on soft interventions in the causal modelling literature (e.g., Eberhardt & Scheines 2007; Korb and Nyberg 2006).

However, the problem of psychological interventions is not solved by allowing for soft interventions. There is a further, equally important reason why interventions on psychological variables are problematic: Psychological interventions typically *change several variables simultaneously*. For example, suppose we wanted to find out whether *pessimistic thoughts* cause *problems in concentration*. In order to do this, we would have to find out what would happen to *problems in concentration* if we were to intervene just on *pessimistic thoughts* without perturbing other psychological states with the intervention. However, how could we intervene on *pessimistic thoughts* without changing, for example, *depressive mood* or *feelings of guilt*? As an actual scientific example, consider a network of psychological variables that includes, among others, the items *alert*, *happy*, and *excited* (Pe et al. 2015). How could we intervene on just one of those variables without changing the others?

One reason why performing "surgical" interventions that only change one psychological variable is so difficult is that there is no straightforward way of manipulating or changing the values of psychological variables (as in, for example, electrical circuits). Interventions in psychology have to be done, for example, through verbal information (as in the example of John above) or through visual/auditory stimuli, and such manipulations are not precise enough to manipulate just one psychological variable. Also state-of-the-art neuroscientific methods such as Transcranial Magnetic Stimulation affect relatively large areas of the brain, and are not suited for intervening on specific psychological variables. Currently, and in the foreseeable future, there is no realistic

way of intervening on a psychological variable without at the same time perturbing some other psychological variables.

Thus, it is likely that most or even all psychological interventions do not just change the target variable X, but also some other variable(s) in the system. In the causal modelling literature, interventions of this kind have been dubbed *fat-handed*[6] interventions (Baumgartner and Gebharter 2016; Eberhardt & Scheines 2007; Scheines 2005). For example, an intervention on pessimistic thoughts that also immediately changes depressive mood is fat-handed. Fat-handed interventions have been recently discussed in philosophy of science, but mainly in the context of mental causation and supervenience (e.g., Baumgartner and Gebharter 2016, Romero 2015), and the fact that psychological interventions are likely to be systematically fat-handed (for reasons unrelated to supervenience) has not yet received attention.

An additional complication is that it is difficult check what a psychological intervention precisely changed, and to what extent it was fat-handed (and soft). In fields such as biology or physics there are usually several independent ways of measuring a variable: for example, temperature can be measured with mercury thermometers or radiation thermometers, and the firing rate of a neuron can be measured with microelectrodes or patch clamps. However, measurements of psychological variables, such as emotions or thoughts, are based on self-reports, and there is no further independent way of verifying that these reports are correct. Moreover, only a limited number of psychological variables can be measured at a given time point, so an intervention may always have unforeseen effects on unmeasured variables.

[6] According to Scheines (2005), this term was coined by Kevin Kelly.

Why are fat-handed interventions so problematic for interventionist causal inference? The reason becomes clear when looking at condition I3: The intervention should not change any variable Z that is on a causal pathway that leads to Y (except, of course, those variables that are on the path between X and Y). If the causal structure of the system under study is known, as well as the changes that the intervention causes, then this condition can sometimes be satisfied even the intervention was fat-handed. However, in the context of intervening on psychological variables, neither the causal structure nor the exact effects of the interventions are known. Thus, when the intervention is fat-handed, it is not known whether I3 is satisfied or not, and in many cases it is likely to be violated. In other words, we cannot assume that the intervention was an unconfounded manipulation of X with respect to Y, and cannot conclude that X is a cause of Y.

4. The Problem of "Holding Fixed"

The next problem that I will discuss is related to the last part of the definition of interventionist causation: X is a cause of Y (in variable set **V**) if and only if it is possible to intervene on X to change Y *when all other variables (in **V**) that are not on the path from X to Y are held fixed to some valu*e. The motivation for this requirement is to make sure that the change in Y is really due to the change X, and not due to some other cause of Y. To a large extent, this is just another way of stating what is already expressed in the definition of an intervention, in conditions I3 and I4: The intervention should not be confounded by any cause of Y that is not on the path between X and Y.[7] In the previous section, we saw that fat-handed interventions pose a challenge for

[7] In recent publications, Woodward often gives a shorter definition of causation that does not include the "holding fixed" part, for example: "X causes Y if and only if under some interventions on X (and possibly other variables) the value of Y changes" (Woodward 2015). This is understandable, as the definition of intervention already contains conditions I3 and I4, which effectively imply holding fixed potential causes of Y that are correlated with the intervention and are not on the path from X to Y. However, there are also good reasons why the full definition has to

satisfying this condition. However, as I will now show, it is problematic in psychology also for more general reasons.

In psychology, it is impossible to hold psychological variables fixed in any concrete way: We cannot "freeze" mental states, or ask an individual to hold her thoughts constant. Thus, the same effect has to be achieved indirectly, and the gold standard for this is Randomized Controlled Trials (RCTs) (Woodward 2003, 2008). RCTs have their origin in medicine, but are widely used in psychology and the social sciences as well (Clarke et al. 2014; Shadish, Cook and Campbell 2002; Shadish and Sullivan 2012). The basic idea of RCTs is to conduct a trial with two groups, the test group and the control group, which are as similar to one another as possible, but the test group receives the experimental manipulation and the control group does not. If the groups are large enough and the randomization is done correctly, any differences between the groups should be only due to the experimental manipulation. If everything goes well, this in effect amounts to "holding fixed" all off-path variables.

However, this methodology has an important limitation that has been overlooked in the literature on interventionism. As the effect of "holding fixed" is based on the difference between the groups as wholes, it only applies at the level of the group, and not at the level of individuals. For this reason, results of RCTs hold for the study population as a whole, but not necessarily for particular individuals in the population (cf. Borsboom 2005, Molenaar & Campbell 2009). For example, if we discover that pessimistic thoughts are causally related to problems of

include the second component as well. For example, consider a situation where we intervene on X with respect to Y, and Y changes, but this change is fully due to a change in variable Z, which is a cause of Y that is *uncorrelated* with the intervention variable. In this situation, without the "holding fixed" requirement we would falsely conclude that X is a cause of Y.

concentration in the population under study, it does not follow that this causal relationship holds in John, Mary, or any other specific individual in the population. This is related to the "fundamental problem of causal inference" (Holland 1986): Each individual in the experiment can belong to only one of the two groups (control or test group), and therefore cannot act as a "control" for herself, so only an average causal effect can be estimated. What this implies for causal inference in psychology is that when a causal relationship is discovered through an RCT, we cannot infer that this relationships holds for any specific individual in the population (see also Illari & Russo 2014, ch. 5).

This does not mean that the population-level findings based on RCTs are uninformative or useless. The point is rather that we currently have no understanding of when, to what extent and under what circumstances they also apply to the individuals in the population. This of course applies also to other fields where RCTs are used, such the biomedical sciences. Indeed, especially in the context of personalized medicine, the fact that RCTs are as such not enough to establish individual-level causal relationships has recently become a matter of discussion (e.g., de Leon 2012).

It might be tempting to simply look at the data more closely and find those individuals for whom the intervention on X actually corresponded with a change in Y. However, it would be a mistake to conclude that in those individuals the change in Y was caused by X. It might very well have been caused by some other cause of Y, as possible confounders were only held fixed at the group level, not at the individual level.[8] Thus, in RCTs possible confounders can only be held fixed at

[8] Would it be possible for a causal relationship to hold at the population level, but not for any individual in the population? Probably not, if the relationship is genuine: Weinberger (2015) has argued that there has to be at least *one* individual in the population for whom the relationships holds. However, in the context of discovery, it is

the group level, and this does not warrant causal inferences that apply to specific individuals. This is further limitation to interventionist causal inference in psychology.

5. Finding psychological causes without interventions

One possible response to the concerns raised in the previous two sections is that interventionism does not require that interventions are actually performed: As briefly mentioned in section 2, what is necessary is to know what *would* happen if we *were* to perform the right kinds of interventions. In other words, in order to establish that X is a cause of Y, it is enough to know that if we *were* to intervene on X with respect to Y (while holding off-path variables fixed), then Y *would* change. For example, it is beyond doubt that the gravitation of the moon causes the tides, even though no one has ever intervened on the gravitation of the moon to see what happens to the tides, and such an intervention would be practically impossible (Woodward 2003). Similarly, it could be argued that even though it is practically impossible to do (ideal) interventions on psychological variables, the knowledge on the effects of interventions could be derived in some other way. Let us thus consider to what extent this could be possible.

The state-of-the-art method for deriving (interventionist) causal knowledge when data on interventions is not available is *Directed Acyclic Graphs (DAGs)*, which were briefly mentioned in the introduction (see also Malinsky & Danks 2018, Pearl 2000, Spirtes, Glymour and Scheines 2000, Spirtes & Zhang 2016). Causal discovery algorithms based on DAGs take purely

possible that a causal *finding* at the population level is just an artefact of heterogeneous causal structures at the individual level, and therefore does not apply to any individual in the population.

observational data as input, and based on conditional independence relations, find the causal graph that best fits the data. In principle, these algorithms can be used for psychological data, with the aim of discovering causal relationships between psychological variables.

However, even though these algorithms do not require experimental data, they do require data from which conditional independence relations can be reliably drawn, and they (implicitly) assume that the variables that are modelled are independently and surgically manipulable (Malinsky & Danks 2018). In contrast, as should be clear from the above discussion, measurements of psychological variables typically come with a great deal of uncertainly, and it is not clear to what extent they can be independently manipulated. Moreover, causal discovery algorithms standardly assume *causal sufficiency*, that is, that there are no unmeasured common causes that could affect the causal structure (Malinsky & Danks 2018; Spirtes & Zhang 2016). The reason for this is that if two or more variables in the variable set have unmeasured common causes, then the inferences concerning the causal relationships between those variables will be either incorrect or inconclusive. However, missing common causes is likely the norm rather than the exception when it comes to psychological variables. For example, if the variable set consists of, say, 16 emotion variables, how likely is it that *all* relevant emotion variables have been included? And even if all emotion variables that are common causes to other emotion variables are included, is it plausible to assume that there are no further cognitive or biological variables that could be common causes to some of the emotion variables? As similar questions can be asked for any context involving psychological variables, causal sufficiency is a very unrealistic assumption for psychological variable sets.

For these reasons, psychological data sets are rather ill-suited for causal discovery algorithms, and these algorithms cannot be treated as reliable guides to interventionist causal knowledge in psychology. It is likely that the problems of psychological interventions discussed in the previous sections are not just practical problems in carrying out interventions, but reflect the immense complexity of the system under study (the human mind-brain), and therefore cannot be circumvented by just using non-experimental data (see, however, section 7 for a different approach).

6. Psychological interventions: A summary

To summarize, what I have argued so far is that interventionist causal inference in psychology faces several obstacles: (1) Psychological interventions are typically *both* fat-handed *and* soft: They change several variables simultaneously, and do not completely determine the value(s) of the variable(s) intervened upon. It is not known to what extent such interventions give leverage for causal inference. (2) Due to the nature psychological measurement, the degree to which a psychological intervention was soft and fat-handed, or more generally, what the intervention in fact did, is difficult to reliably estimate. (3) Holding fixed possible confounders is only possible at the population level, not at the individual level, and it is not known under what conditions population-level causal relationships also apply to individuals. (4) Causal inference based on data without interventions requires assumptions that are unrealistic for psychological variable sets. Taken together, these issues amount to a formidable challenge for finding psychological causes.[9]

[9] Baumgartner (2009, 2012, 2018) has argued that mental-to-physical supervenience makes it impossible to satisfy the Woodwardian conditions on interventions, and that if interventionism is modified to accommodate supervenience relationships (as in Woodward 2015), the result is that any causal structure with a psychological

7. Discussion

Although various metaphysical and conceptual issues related to psychological causation have been extensively discussed in philosophy of science, little attention has been paid to the *discovery* of psychological causes. In this paper, I have contributed to filling this lacuna, by discussing the search for psychological causes in the framework of the interventionist theory of causation. The upshot is that finding individual psychological causes faces daunting challenges. The problems in holding fixed confounders and performing interventions need to be taken into account when trying to establish a psychological causal relationship, or when making claims about psychological causes.

However, I do not want to argue that finding psychological causes is *impossible*, or that researchers should stop looking for psychological causes. Rather, my aim is to contribute to getting a better understanding of the limits of finding causes in psychology, and the challenges involved. This can also lead to positive insights regarding causal inference in psychology. One such insight is that more attention should be paid to *robust inference* or *triangulation*. Often when individual methods or sources of evidence are insufficient or unreliable, as is the case here, what is needed is a more holistic approach. A widespread (though not uncontroversial) idea in philosophy of science is that evidence from several independent sources can lead to a degree of confidence even if the sources are individually fallible and insufficient (Eronen 2015, Kuorikoski

cause becomes empirically indistinguishable from a corresponding structure where the psychological variable is epiphenomenal. If this reasoning is correct, it leads to a further (albeit more theoretical) problem for interventionist causal inference: Any empirical evidence for a causal relationships with a psychological cause is equally strong evidence for a corresponding epiphenomenal structure, and it is not clear which structure should be preferred and on what grounds.

& Marchionni 2017, Munafo & Smith 2017, Wimsatt 1981, 1994/2007). For example, there is no single method or source of evidence that would be individually sufficient to establish that the anthropogenic increase in carbon dioxide is the cause for the rise in global temperature, but there is so much converging evidence from many independent sources that scientists are confident that this causal relationship exists. Similarly, evidence for a psychological causal relationship could be gathered from many independent sources: Several different (soft and fat-handed) interventions involving different variables, multilevel models based on time-series data, single-case observational studies, and so on.[10] If they all point towards the same causal relationships, this may lead to a degree of confidence in the reality of that relationship. However, how this integration of evidence would exactly work, and whether it can actually lead to sufficient evidence for psychological causal relationships, are open questions.

A related point is that psychological research can also make substantive progress *without* establishing causal relationships. Often important discoveries in psychology have not been discoveries of causal relationships, but rather discoveries of robust *patterns* or *phenomena* (Haig 2012, Rozin 2001, Tabb and Schaffner 2017). Consider, for example, the celebrated discovery that people often do not reason logically when making statistical predictions, but rely on shortcuts, for example, grossly overestimating the likelihood of dying in an earthquake or terror attack (Kahneman & Tversky 1973). In other words, when we reason statistically, we often rely on heuristics that lead to biases. The discovery of this phenomenon had nothing to with methods of causal inference (Kahneman and Tversky 1973), and its significance is not captured by describing causal relationships between variables. In fact, the causal mechanisms underlying the

[10] See also Peters, Bühlmann, & Meinshausen (2016), who present a formal model for inferring causal relationships based on their stability under different kinds of (non-ideal) interventions.

heuristics and biases of reasoning are still unknown. Similar examples abound in psychology: Consider, for example, groupthink or inattentional blindness. Of course, there are likely to be causal mechanisms that give rise to these phenomena, but the phenomena are highly relevant for theory and practice even when we know little or nothing about those underlying mechanisms (which is the current situation). This, in combination with the challenges discussed in this paper, suggests that (philosophy of) psychology might benefit from reconsidering the idea that discovering causal relationships is central for making progress in psychology.

Finally, one might wonder whether the problems I have discussed here are restricted to just psychology. Indeed, I believe that the arguments I have presented are more general, and apply to any other fields where there are similar problems with soft and fat-handed interventions and controlling for confounders. There is probably a continuum, where psychology is close to one end of the continuum, and at the other end we have fields where ideal interventions can be straightforwardly performed and variables can be easily held fixed, such as engineering science. Fields such as economics and political science are probably close to where psychology is, as they also face deep problems in making (ideal) interventions and measuring their effects. Same holds for neuroscience, at least cognitive neuroscience: The problems of soft and fat-handed interventions and holding variables fixed apply just as well to brain areas as to psychological variables (see also Northcott forthcoming). Thus, appreciating the challenges I have discussed here and considering possible reactions to them could also benefit many other fields besides psychology.

To conclude, I have argued in this paper that there are several serious obstacles to the discovery of psychological causes. As it is widely assumed in both psychology and its philosophy that the discovery of causes is a central goal, these obstacles need to be explicitly discussed, taken into account, and studied further.

References

Baumgartner, M. (2013). Rendering Interventionism and Non-Reductive Physicalism Compatible. *dialectica* 67: 1-27.

Baumgartner, M. (2018). The Inherent Empirical Underdetermination of Mental Causation. *Australasian Journal of Philosophy*.

Baumgartner, M and Gebharter, A. (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science* 67: 731-756.

Borsboom, Dennny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press

Borsboom, Denny and Anqelique O. Cramer. 2013. "Network analysis: an integrative approach to the structure of psychopathology." *Annual review of clinical psychology* 9: 91-121.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203.

Campbell, John. 2007. "An interventionist approach to causation in psychology." In: A. Gopnik & L. Schulz (eds.) *Causal Learning. Psychology, Philosophy, and Computation*. Oxford: Oxford University Press, 58–66.

Chirimuuta, Mazviita. Forthcoming. “Explanation in Computational Neuroscience: Causal and Non-causal.” *British Journal for the Philosophy of Science*. DOI:https://doi.org/10.1093/bjps/axw034

Clarke, B., Gillies, D., Illari, P, Russo, F., & Williamson, J. 2014. “Mechanisms and the evidence hierarchy.” *Topoi* 33: 339-360.

de Leon, J. (2012). Evidence-based medicine versus personalized medicine: are they enemies? *Journal of clinical psychopharmacology*, *32*(2), 153-164.

Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science*, *80*(5), 684-696.

Eberhardt, F. (2014). Direct causes and the trouble with soft interventions. *Erkenntnis*, *79*(4), 755-777.

Eberhardt, Frederick and Richard Scheines. 2007. “Interventions and causal inference.” *Philosophy of Science* 74: 981–995.

Eronen, Markus. Forthcoming. “Interventionism for the Intentional Stance: True Believers and Their Brains.” *Topoi*.

Hamaker, Ellen L. 2011. “Why researchers should think **“**within-person.**””** In M. R. Mehl, & T. A. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Kahneman, Daniel and Amos Tversky. 1973. “On the psychology of prediction.” *Psychological Review* 80: 237-251.

Kendler, Kenneth S. and John Campbell. 2009. Interventionist causal models in psychiatry: repositioning the mind-body problem. *Psychological Medicine* 39: 881-887.

Korb, K. B., & Nyberg, E. 2006. "The power of intervention." *Minds and Machines* 16*:* 289-302.

Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science*, *83*, 227-247.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470.

Menzies, Peter. 2008. "The exclusion problem, the determination relation, and contrastive causation." In J. Hohwy & J. Kallestrup (Eds.) *Being Reduced* (pp. 196-217). Oxford: Oxford University Press.

Molenaar, Peter and Cynthia Campbell. 2009. "The new person-specific paradigm in psychology." *Current Directions in Psychological Science* 18: 112-117.

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature* 553, 399-401

Northcott, R. (forthcoming). Free will is not a testable hypothesis. *Erkenntnis*.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... & Kuppens, P. 2015. "Emotion-network density in major depressive disorder." *Clinical Psychological Science*, *3*(2), 292-300.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, Judea. 2009. "Causal inference in statistics: An overview." *Statistics surveys* 3: 96-146.

Pearl, Judea. 2014. "Comment: understanding simpson's paradox." *The American Statistician* 68: 8-13.

Peters, J. , Bühlmann, P. and Meinshausen, N. (2016), Causal inference by using invariant prediction: identification and confidence intervals. J. R. Stat. Soc. B, 78: 947-1012. doi:10.1111/rssb.12167

Rescorla, Michael. Forthcoming. “An interventionist approach to psychological explanation.” *Synthese*.

Reutlinger, Alexander and Juha Saatsi (eds.). 2017. *Explanation Beyond Causation*. Oxford: Oxford University Press.

Romero, F. (2015). Why there isn’t inter-level causation in mechanisms. *Synthese*, *192*(11), 3731-3755.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2-14.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science, 72*(5), 927-940.

Shadish W. R., Cook T. D. and Campbell D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin; Boston.

Shadish, W. R., & Sullivan, K. J. 2012. “Theories of causation in psychological science.” In H. Cooper et al. (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 23-52). Washington, DC: American Psychological Association.

Shapiro, Lawrence. 2010. “Lessons from causal exclusion.” *Philosophy and Phenomenological Research*, 81, 594-604.

Shapiro, Lawrence. 2012. “Mental manipulations and the problem of causal exclusion.” *Australasian Journal of Philosophy*, 90, 507–524.

Shapiro, Lawrence and Elliott Sober. 2007. “Epiphenomenalism: the dos and the don’ts.” In G. Wolters & P. Machamer (Eds.) *Thinking about causes: from Greek philosophy to modern physics* (pp. 235–264). Pittsburgh, PA: University of Pittsburgh Press.

Spirtes, Peter, Glymour, Clark and Richard Scheines. 2000. *Causation, prediction, and search.* New York: Springer.

Tabb, K., & Schaffner, K. F. (2017). Causal pathways, random walks and tortuous paths: Moving from the descriptive to the etiological in psychiatry. In: Kendler, K. S., & Parnas, J. (Eds.) *Philosophical Issues in Psychiatry IV: Nosology* (pp. 342-360). Oxford: Oxford University Press.

Weinberger, Naftali. 2015. “If intelligence is a cause, it is a within-subjects cause.” *Theory & Psychology*, 25(3), 346-361.

Woodward, James. 2003. *Making things happen. A theory of causal explanation.* Oxford: Oxford University Press.

Woodward, James. 2008. “Mental causation and neural mechanisms.” In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: new essays on reduction, explanation, and causation.* Oxford: Oxford University Press: 218-262

Woodward, James. 2015a. “Interventionism and causal exclusion.” *Philosophy and Phenomenological Research* 91, 303-347.

Woodward, James. 2015b. “Methodology, ontology, and interventionism.” *Synthese* 192, 3577-3599.

Woodward, James & Christopher Hitchcock. 2003. “Explanatory Generalizations, Part I: A Counterfactual Account.” *Noûs* 37(1): 1–24.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

Why Replication is Overrated

Current debates about the replication crisis in psychology take it for granted that direct replication is valuable and focus their attention on questionable research practices in regard to statistical analyses. This paper takes a broader look at the notion of replication as such. It is argued that all experimentation/replication involves individuation judgments and that research in experimental psychology frequently turns on probing the adequacy of such judgments. In this vein, I highlight the ubiquity of conceptual and material questions in research, and I argue that replication is not as central to psychological research as it is sometimes taken to be.

1. Introduction: **The "Replication Crisis"**

In the current debate about replicability in psychology, we can distinguish between (1) the question of why not more replication studies are done (e.g., Romero 2017) and (2) the question of why a significant portion (more than 60%) of studies, when they *are* done, fail to replicate (I take this number from the Open Science Collaboration, 2015). Debates about these questions have been dominated by two assumptions, namely, first, that it is in general desirable that scientists conduct replication studies that come as close as possible to the original, and second, that the low replication rate can often be attributed to statistical problems with many initial studies, sometimes referred to as "p-hacking" and "data-massaging."[1]

[1] An important player in this regard is the statistician Andrew Gelman who has been using his blog as a public platform to debate methodological problems with mainstream social psychology (http://andrewgelman.com/).

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

I do not wish to question that close (or "direct") replications can sometimes be epistemically fruitful. Nor do I wish to question the finding that are severe problems in the statistical analyses of many psychological experiments. However, I contend that the focus on formal problems in data analyses has come at the expense of questions about the notion of *replication* as such. In this paper I hope to remedy this situation, highlighting in particular the implications of the fact that psychological experiments in general are infused with conceptual and material presuppositions. I will argue that once we gain a better understanding of what this entails with respect to replication, we get a deeper appreciation of philosophical issues that arise in the investigative practices of psychology. Among other things, I will show that replication is not as central to these practices as it is often made out to be.

The paper has three parts. In part 1 I will briefly review some philosophical arguments as to why there can be no exact replications and, hence, why attempts to replicate always involve individuation judgments. Part 2 will address a distinction that is currently being debated in the literature, i.e., that between direct and conceptual replication, highlighting problems and limitations of both. Part 3, finally, will argue that a significant part of experimental research in psychology is geared toward exploring the shape of specific phenomena or effects, and that the type of experimentation we encounter there is not well described as either direct or conceptual replication.

2. The Replication Crisis and the Ineliminability of Concepts

When scientists and philosophers talk about successfully replicating an experiment, they typically mean that they performed the same experimental operations/interventions. But what does it mean to perform "the same" operations as the ones performed by a previous experiment? With regard to this question, I take it to be trivially true that two experiments cannot be identical: At the very least, the time variable will differ. Replication can therefore at best aim for *similarity* (Shavit & Ellison 2017), as is also recognized by some authors in psychology. In this vein, Lynch et al (2015) write that "[e]xact

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

replication is impossible" (Lynch et al 2015, 2), arguing that at most advocates of direct replication can aim for is to get "as close as possible," i.e., to conduct an experiment that is similar to the previous one. In the literature, such experiments are also referred to as "direct replications." (e.g., Pashler & Harris 2912).[2]

The notion of similarity is, of course, also notoriously problematic (e.g., Goodman 1955), since any assertion of similarity between A and B has to specify with regard to what they are similar. In the context of experimentation, the relevant kinds of specifications already presuppose conceptual and material assumptions, many of which are not explicated, about the kinds of factors one is going to treat as relevant to the subject matter (see also Collins 1985, chapter 2). Such conceptual decisions will inform what one takes to be the "experimental result" down the line (Feest 2016). For example, If I am interested in whether listening to Mozart has a positive effect on children's IQ, I will design an experiment, which involves a piece by Mozart as the independent variable and the result of a standardized IQ-test at a later point. Now if I get an effect, and if I call it a Mozart effect, I am thereby assuming that the piece of music I used was causally responsible *qua being a piece by Mozart*. Moreover, when I claim that it's an effect on intelligence, I am assuming that the test I used at the end of the experiment *in fact measured intelligence*. These judgments rely on conceptual assumptions already built into the experiment qua choice of independent and dependent variables. In addition, I need *material assumptions* to the effect that potentially confounding variables have been controlled for. I take this example to show that whenever we investigate an effect *under a description*, we cannot avoid making conceptual assumptions when determining whether an experiment has succeeded or failed. This goes for original experiments as well as for replications.

[2] Both advocates and critics of direct replication sometimes contrast such replications with "conceptual" replications" (Lynch et al 2015). We will return to this distinction below.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

One obvious rejoinder to this claim might be to say that replication attempts need not investigate effects under a description. They might simply imitate what the original experiment did, with no particular commitment to what is being manipulated or measured. But even if direct replications need not explicitly replicate effects under a description, I argue that they nonetheless have to make what Lena Soler calls "individuation judgments" (Soler 2011). For example, the judgment that experiment 2 is relevantly similar to experiment 1 involves the judgment that experiment 2 does not introduce any confounding factors that were absent in experiment 1. However, such judgments have to rely on some assumptions about what is relevant and what is irrelevant to the experiment, where these assumptions are often unstated auxiliaries. For example, I may (correctly or incorrectly) tacitly assume that temperature in the lab is irrelevant and hence ignore this variable in my replication attempt.

It is important to recognize that the individuation judgments made in experiments have a high degree of epistemic uncertainty. Specifically, I want to highlight what I call the problem of "conceptual scope," which arises from the question of how the respective independent and dependent variables are described. Take, for example, the above case where I play a specific piece by Mozart in a major key at a fast pace. A lot hangs on what I take to be the relevant feature of this stimulus: the fact that it's a piece by Mozart, the fact that it's in a major key, the fact that it's fast? etc. Depending on how I describe the stimulus, I might have different intuitions about possible confounders to pay attention to. For example, if I take the fact that a piece is by Mozart as the relevant feature of the independent variable, I might control for familiarity with Mozart. If I take the relevant feature to be the key, I might control for mood. Crucially, even though scientists make decisions on the basis of (implicit or explicit) assumptions about conceptual scope, their epistemic situation is typically such that they don't know what is the "correct" scope. This highlights a feature of psychological experiments that is rarely discussed in the literature about the replication crisis, i.e., the deep epistemic uncertainty and conceptual openness of much

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

research. This concerns both the initial and the replication study. Thus, concepts are ineliminable in experimental research, while at the same time being highly indeterminate.

3. Is the dichotomy between direct and less direct replication pragmatically useful?

One way of paraphrasing what was said above is that all experiments involve individuation judgments and that this concerns both original and replication studies. While this serves as a warning against a naïve reliance on direct (qua non-conceptual) replication, it might be objected that direct replications nonetheless make unique epistemic contributions. This is indeed claimed by advocates of both direct and less direct (="conceptual") replication alike. I will now evaluate claims that have aligned the distinction between direct and "conceptual" with some relevant distinctions in scientific practice, such as that between the aim of establishing the existence of a phenomenon and that of generalizing from such an existence claim on the one and that between reliability and validity on the other. I will argue that while these distinctions are heuristically useful, but on closer inspection bring to the fore exactly the epistemological issues just discussed.

3.1 Existence vs. Generalizability

Many scientists take it as given that there cannot be two identical experiments, but nonetheless argue that there is significant epistemic merit in trying to get *close enough*., i.e., to conduct direct replications. In turn, the notion of a direct replication is frequently contrasted with that of a "conceptual" replication. In a nutshell, direct replications essentially try to redo "the same" experiment (or at least something very close), whereas the conceptual replications try to operationalize the same question or concept/effect in a different way. The advantage of direct replications, as viewed by its advocates, is that by being able to redo an experiment faithfully and to create the same effect, one can show that the effect was real: "Exact and very close replications establish the basic existence and stability of a

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

phenomenon by falsifying the (null) hypothesis that observations simply reflect random noise" (LeBel et al, forthcoming, 7).

Advocates of conceptual replication don't deny this advantage of close replications, but hold that we want more than to establish that a given effect – created under very specific experimental conditions – is real. We want to know whether our findings about it can be generalized to: "When the goal is generalization, we argue that 'imperfect' conceptual replications that stretch the domain of research may be more useful" (Lynch et al 2015, 2). From a strictly Popperian perspective, the idea that non-falsification of the hypothesis of random error can provide proof of stability and existence is questionable , of course. But even if we abandon Popperian ideology here and take the falsification of H0 (that the initial effect was due to random error) to point to the truth of H1 (that there is a stable effect), the question is how to describe the effect. In other words, when claiming to have confirmed an effect, we have to say *what kind of effect* it is. And there we face the following dilemma:

a) Either we describe the effect as highly specific to very local experimental circumstances, involving the choice of a specific independent variable, delivered in a specific way etc.

b) Or we describe it in slightly broader terms, e.g., as a Mozart effect.

Advocates of direct replication might indeed endorse something like a), thereby exhibiting the kind of caution that motivated early operationists, in that no claim is made beyond the confines of a specific experiment. If, on the other hand, psychologists endorsed a description such as b), they would immediately run into the question of conceptual scope, i.e., the question *under what description* the independent variable can be said to have caused an effect. I argue that no amount of direct replication can answer this question, and hence, even if direct replication can confirm the existence of an effect, it cannot say what kind of effect. By asserting this, I am not saying that it's never useful to do a direct replication. My claim is merely that it will tell us relatively little. More pointedly: Direct replication can (perhaps) provide evidence for the existence of something, but it cannot say *existence of what*. Rolf

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

Zwaan makes a similar point when he states that "replication studies "tell us about the reliability of those findings. They don't tell us much about their validity." (Zwaan 2013).

In a similar vein, I argue that direct replication, with its narrow focus on ruling out random error, is epistemically unproductive, because it has nothing to say about *systematic error*. Systematic error arises if one erroneously attributes an effect to a specific feature of the experiment, when it is in fact due to another feature of the experiment. This can include, but is not limited to, the above-mentioned problem of conceptual scope. Fiedler et al. (2012) make a similar point when they argue that a narrow focus on falsification (with the aim of avoiding false positives) can be detrimental to the research process. Differently put, by privileging direct replication, we are not in a position to inquire about the kind of effect in question. This question, I argue, is best addressed by paying close attention to the possibility of systematic error, and hence by doing conceptual work. In other words, experimentally probing into systematic errors of conceptual scope is a valuable and productive part of the research process as it enables scientists to gradually explore what kind of effect (if any) they are looking at.[3]

3.2 Generality

I have argued that (a) scientists typically produce effects under a description and (b) that it can be epistemically productive to probe the scope of the description and to investigate the possibility of systematic error with regard to experiments that draw on such descriptions. It is epistemically productive, because it forces scientists to explore the nature and boundaries of the effect they are investigating. With this I have argued against a narrow focus on direct replication and I have cautioned against overstating the epistemic merits of such replication. But when we are concerned with effects

[3] I take this to be a contribution to arguments that philosophers of experimentation have made for a long time; e.g., Mayo 1996.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

under a description, we are confronted with questions about the adequacy of the description. It is this question that advocates of "conceptual replication" claim to be able to address when they emphasize that their approach can deliver generality (over mere existence).

We have to distinguish between two notions of generality, namely (a) what kinds of descriptions one can generalize or infer to *within the experiment*, and (b) does the effect in question hold *outside the lab* (see Feest & Steinle 2016). These types of generality are also sometimes referred to as internal vs. external validity, respectively (Campbell & Stanley 1966; Guala 2012), where the former refers to the quality of inferences within an experiment and the latter refers to the quality of inferences from a lab to the world. The notion of generalizability raises questions about two kinds of validity. My focus here will be on internal validity, i.e., with the question of whether the effect generated in an experiment really exists as described by the scientist.[4]

Internal validity can fail to hold because of epistemic uncertainties regarding confounding variables both internal and external to experimental subjects. For example, prior musical training might make a difference to how one responds to Mozart music, but the experimenter may not have taken this into consideration in their design. But internal validity can also fail to hold is by virtue of what I have referred to as the problem of conceptual scope (for example, we may refer to the effect as a Mozart effect when it is in fact a Major-key effect). Effectively, when I treat a major-key effect as a Mozart effect, I have misidentified the relevant causal feature of the stimulus. In turn, this means that I will neglect to control for major/minor key as I will regard this as irrelevant, which can result in systematic errors. In both cases, scientists can go wrong in their individuation judgment. What is at stake is not whether there is an effect, but what kind of effect it is. Now, given that those kinds of problems can

[4] In this respect I differ from some advocates of conceptual replication, who have highlighted external validity as a desideratum (E.g., Lynch 1982, 3/4).

occur, we turn to the question of whether "conceptual replication" has an answer. I will now argue that it does not.

To explain this, let me return to the above characterization of conceptual replication, according to which such replication consists in repeating an experiment, using different operationalizations of the same construct. For example, a conceptual replication of an experiment about the Mozart effect might operationalize the concept Mozart effect differently by using a different piece of Mozart music and/or a different measure of spatial reasoning. But there is a major caveat here: If I want to compare the results of two experiments that operationalized the same construct differently, I already have to presuppose that both operationalizations in fact have the same conceptual scope, i.e., that they in fact individuate the same effect. But this would be begging the question, since after all – given the epistemic uncertainty and conceptual openness highlighted above – that's precisely what's at issue. Differently put, experiment 2 might or might not achieve the same result as experiment 1, but the reason for this would be underdetermined by the experimental data. Thus, the problem of conceptual scope prevents us from being able to say whether we have succeeded in our conceptual replication.

Given the uncertainties as to whether one has in fact succeeded in conceptually replicating a given experiment, I am weary of the language of replication here. If anything, I would argue that the method in question should be regarded as a research strategy that is aimed at helping to demarcate and explore the very subject matter under investigation. But as I will argue now, this is perhaps better described as exploration, not as replication.

4. Putting Replication in its Proper Place

The conclusion of the previous paragraphs seems pretty bleak: Direct replication is either extremely narrow in what it can deliver or it runs into the joint problems of confounders and conceptual scope. Conceptual replication, on the other hand, cannot come to the rescue, because it also runs into the

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

exact same problems. Should we then throw up our hands and conclude that since ultimately neither direct nor conceptual replication are possible the crisis of replication is much more severe than we previously thought? This would be the wrong conclusion, however. This would only follow if replication was in fact as central to research as it is sometimes taken to be. I claim that it is not. My argument for these claims has three parts. The first part holds that exploring (the possibility of) systematic errors is an important part of the investigative process, which is not well described as replication. Second, if we take seriously this process of exploring and delineating the relevant phenomena, we find that there is indeed a great deal of uncertainty in psychological research, but this, in and of itself, does not necessarily constitute a crisis. Lastly, while it is fair to say that there is a crisis of confidence in current psychology, it is not well described as a replication crisis.

Let me begin with the first point. I have argued that direct replication (even where it is successful) is of limited value, because it can at most rule out random error, but completely fails to be able to address systematic error. But if we appreciate (as I have argued we should) that direct replication inevitably involves individuation judgments, it is obvious that there is always a danger of systematic error, because I have to assume that all confounding variables have been controlled for. One important class of confounders follows from what I have referred to as the problem of conceptual scope, i.e., the difficulty of correctly describing both the independent variable responsible for a given effect and the dependent variable.[5] Epistemically productive experimental work, I claim, therefore needs to focus on systematic errors, specifically those brought about by unstated auxiliary assumptions.

Indeed, if we look at the story of the Mozart effect, we find that this is exactly what happened. This example also nicely illustrates my claim about the conceptual openness and epistemic uncertainty

[5] My focus here has been mainly on the former. But of course the problem of conceptual scope concerns both.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

in many areas of experimental psychology. The Mozart effect was first posited by Rauscher and colleagues (Rauscher et a. 1993). It can now be regarded as largely debunked. While it is true that several people tried (and failed) to replicate the effect (e.g., Newman et al. 1995; Steele 1997), it is important to look at the details here. It is not the case that the effect was simply abandoned for lack of replicability. Rather, when we look at the back and forth between Rauscher and her critics, we find that the discussion turned on the choices and interpretations of independent and dependent variables. In this vein, Newman et al (1995) and Steele (1997) used different dependent variables, prompting Rauscher to argue that her effect was more narrowly confined to the kind of spatial reasoning measured by the Stanford-Binet. I suggest that we interpret this case as one where Rauscher was forced to confront (and retract) an unstated auxiliary assumption of her initial study, namely that the spatial reasoning subtest of the Stanford-Binet (which she had used as her dependent variable), was representative of spatial reasoning more generally. Likewise, her choice of the Mozart's Sonata for Two Pianos in D-major as the independent variable was put under considerable pressure by critics, who suggested that the relevant feature of the independent variable was not that it was a piece by Mozart, but that it was up-beat and put subjects in a good mood (Chabris 1999). My point here is that the debates surrounding the Mozart effect are best described as conceptual work, exploring consequences of possible errors that might have arisen from the problem of conceptual scope. At issue, I claim, was not primarily whether Rauscher really found an effect, but rather what was the scope of the effect.

I argue that this is a typical case. Rather than, or in addition to, attempting to conduct direct replications of previous experiments, researchers critically probed some hidden assumptions built into the design and interpretation of the initial experiment. My point here is both descriptive and normative. Thus, I argue that this is a productive way to proceed. However, I claim that it is not well described as replication, let alone conceptual replication. Rather, what we see here is a case in which scientists explore the empirical contours of a purported effect in the face of a high degree of epistemic

uncertainty and conceptual openness, and this is precisely why the case is not well described as employing conceptual replication. The reason for this is quite simple: For a conceptual replication to occur, one needs to already be in the possession of some well-formed concepts, such that they can be operationalized in different ways. It also presupposes that in general the domain is well-understood, such that operationalizations can be implemented and confounding variables can be controlled. But this completely misses the point that researchers often investigate effects precisely because they don't have a good understanding (and hence concept) of what it is.

Therefore I argue that while direct replication can only contribute a very small part to the research process, conceptual replication cannot make up for the shortcomings of direct replication. Instead, productive research should (and frequently does) proceed by exploring, and experimentally testing, hypotheses about possible systematic errors in experiment. Such research, I suggest, can contribute to conceptual development by helping to explore and fine-tune the shape and scope of proposed or existing concepts. The fact that this is riddled with problems does not in and of itself constitute a crisis, let alone a replication crisis.

5. Conclusion

The upshot of the above is that when we talk about the importance of replication, we need to be clear on what we mean by replication and why it is so important, precisely.

In this paper I have argued that if by replication we mean either "direct" or "conceptual" replication, we need to first of all be clear that direct replications are not non-conceptual. I then turned to some alleged epistemic merits of direct replication, for example that they can establish the existence of effects or the reliability of procedures that detect effects. I argued that insofar as such replications involve concepts, they run (among other things) into the problem of conceptual scope, i.e., the difficulty of determining, on the basis of independent and dependent variables of experiments what precisely is

the scope of the effect one is trying to replicate. I highlighted that this is a real and pernicious problem in experimental research in psychology, due to the high degree of epistemic uncertainty and conceptual openness of many fields of research.

While my emphasis of the conceptual nature of replication may suggest that I would be more favorably inclined toward conceptual replication, I have argued that conceptual replication runs into the same problems, and for similar reasons: The very judgement that one has successfully performed a conceptual replication of a previous experiment presupposes what is ultimately the aim of the research, namely to arrive at a robust understanding of the relevant area of research. This, I argue that since conceptual replication presupposes a relatively good grasp of the relevant concepts, it is begging the question, and I suggested instead that researchers (should) engage in a process of specifically investigating possible systematic errors in original studies as a means to develop the relevant concepts. This process is not best described as one of replication, however. Summing up, then, I conclude that in general, replications are less useful and important than is widely assumed – at least in the kind of psychological research I have focused on in this paper.

Now, in conclusion let me return to the notion of a crisis in psychology as it is currently discussed in the literature. Obviously, I do not mean to deny that there is a crisis of confidence in (social) psychology (Earp & Trafimov 2015) as well as in other areas of study. However, based on the analysis provided in this paper, I argue that this crisis is not well described as a crisis of replication. Rather, it seems to be to a large degree a crisis that turns on questionable research practices with regard to the use of statistical methods in psychology (see Gelman & Loken 2014). While acknowledging the valuable philosophical and scientific work that is being done in this area, I suggest that a broader focus on the notion of replication provides us with a deeper appreciation of the conceptual dynamics characteristic of experimental practice.

REFERENCES

Campbell, D. T., and Stanley, J. C. (1966), Experimental and Quasi-Experimental Designs for Research (Chicago: Rand McNally).

Chabris, C. (1999): Prelude or requiem for the 'Mozart Effect?' "Scientific Correspondence", Nature, 4000, 826.

Collins, H. (1985). Changing order. Replication and induction in scientific practice. Chicago and London: The University of Chicago Press.

Earp, Brian & Trafimow, David (2015): "Replication, falsification, and the crisis of confidence in social psychology." Front. Psychol, 19 May 2015 | https://doi.org/10.3389/fpsyg.2015.00621

Feest, U., 2016, "The Experimenters' Regress Reconsidered: Tacit Knowledge, Skepticism, and the Dynamics of Knowledge Generation". Studies in History and Philosophy of Science, Part A 58 34-45.

Feest, U. & Steinle, F., 2016, "Experiment." In P. Humphreys (Ed.): Oxford Handbook of Philosophy of Science. Oxford University Press, 274–295.

Fiedler, K.; Kutzner, F. & Krueger, J. (2012): „The Long Way from alpha-error control to validity proper: Problems with a short-sighted false-positive debate." Perspectives on Psychological Science 7(6), 661-669

Gelman, Andrew & Loken, Eric (2014): The Statistical Crisis in Science. Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up. American Scientist 102 (6) 460-464. DOI: 10.1511/2014.111.460

Goodman, Nelson (1983/1955): Fact. Fiction and Forecast. Harvard University Press; 4 Revised edition edition

Guala, F. (2012), "Philosophy of Experimental Economics." In U. Mäki (ed.), Handbook of the philosophy of science. Vol. 13: Philosophy of Economics (Boston: Elsevier/Academic Press), 597–640

LeBel, E.P.; Berger, D., Campbell, L.; Loving, T. (2017): "Falisifiability is not Optional." Journal of Personality and Social Psychology (forthcoming)

Lynch, J. (1982): "On the External Validity of Experiments in Consumer Research. Journal of Consumer Research 9, 225-239. (December)

Lynch, J.; Bradlow, E.; Huber, J.; Lehmann, D. (2015): "Reflections on the replication corner: In praise of conceptual replication." IJRM ???

Mayo, Deborah (1996): Error and the Growth of Experimental Knowledge. University of Chicago Press.

Newman, J., Rosenbach, J., Burns, K.; Latimer, B., Matocha, H., Vogt, E. (1995: An experimental test of the 'Mozart Effect': Does listening to Mozart improve spatial ability? Perceptual and Motor Skills, 81, 1379-1387.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349

Pashler, Harold & Harris, Christine (2012): "Is the Replication Crisis Overblown?" Perspectives on Psychological Science 7(6), 531-536.

Rauscher, F., Shaw, G.; Ky, K. (1993). Music and spatial task performance. Nature ,365, 611.

Romero, Felipe (2017): "Novelty vs. Replicability. Virtues and Vices in the Reward System of Science." Philosophy of Science.

Shavit, Ayelet & Ellison, Aaron (eds.) (2017): Stepping in the Same River Twice. Replication in Biological Research. Yale University Press

Soler, Lena (2011): "Tacit Elements of Experimental Practices: analytical tools and epistemological consequences." European Journal for Philosophy of Science 1, 393-433.

Steele, K., (2000). Arousal and mood factors in the 'Mozart effect'. Perceptual and Motor Skills, 91, 188-190.

Zwaan, Rolf (2013): "How Valid are our Replication Attempts?"

https://rolfzwaan.blogspot.de/2013/06/how-valid-are-our-replication-attempts.html

Author: Paul L. Franco, UW-Seattle, Department of Philosophy

Contact: pfranco@uw.edu

Title: Speech Act Theory and the Multiple Aims of Science

Abstract: I draw upon speech act theory to understand the speech acts appropriate to the multiple aims of scientific practice and the role of nonepistemic values in evaluating speech acts made relative to those aims. First, I look at work that distinguishes explaining from describing within scientific practices. I then argue speech act theory provides a framework to make sense of how explaining, describing, and other acts have different felicity conditions. Finally, I argue that if explaining aims to convey understanding to particular audiences rather than describe literally across all contexts, then evaluating explanatory acts directed to the public or policymakers involves asking nonepistemic questions.

(Accepted with minor revisions to the PSA 2018 proceedings issue of Philosophy of Science | Revisions not yet made; final version due January 2019)

1. Introduction

Hasok Chang "[complains] about…our [i.e., philosophers of science] habit of focusing on descriptive statements that are either products or presuppositions of scientific work, and our commitment to solving problems by investigating the logical relationships between these statements" (2014, 67–8). He argues philosophers of science should adopt "a change of focus from propositions to actions" (67). Chang suggests, "When we do pay attention to words, it would be better to remember to think of 'how to do things with words', to recall J. L. Austin's (1962) famous phrase" (68).

In this paper, I take Chang's suggestion and argue that attending to Austin's account of the things we do with words can help us understand the multiple goals of scientific practices, the speech acts appropriate to those goals, and the roles of nonepistemic values in evaluating speech acts made relative to those aims. In §2, I give an overview of a few philosophers of science working on explanation who have shifted focus from propositions to explaining.[1] I also briefly relate this work to a few themes in speech act theory. In §3, I give more details of Austin's framework to highlight ways of evaluating speech acts beyond truth and falsity. In §4, I explore the multiple goals of scientific practice, especially goals related to conveying understanding to the general public and policymakers, and the speech acts appropriate to those goals.

2. The things scientists do with words

2.1 Explaining

Consider some recent and not-so-recent work on scientific explanation. Andrea Woody's defense of a functional perspective on explanation aims to motivate "a shift in focus away from explanations, as achievements, toward explaining, as a coordinated activity of communities" (2015, 80). In a similar spirit, Angela Potochnik argues that when looking at explanation, "sidelining the communicative purposes to which explanations are put is a mistake" (2016, 724). She emphasizes that explaining is a communicative act involving a speaker and audience made against a background that shapes the explanations offered. In so

[1] I make no claims Chang influenced the work I canvas.

arguing, Potochnik deliberately recalls Peter Achinstein's claim, "Explaining is an illocutionary act," i.e., a speech act uttered by a speaker with a certain force and for a certain point (1977, 1).

These accounts share in common an emphasis on the importance of the aims of the speaker and audience, and thus the context of utterance in evaluating, to borrow terminology from Austin, the felicity conditions of explanatory speech acts. In particular, we might focus on the aims of the speaker and their audience in requesting and giving explanations, the time and location of an explaining speech act, and, following Woody, "what role(s) [explanations] might play in practice" (2015, 81). In focusing on the explaining act rather than the supposedly stable propositional content of an act of explanation, our attention is drawn to dimensions of evaluation beyond truth and falsity.

On this last point, Nancy Cartwright argues that the functions of a scientific theory to "tell us…what is true in nature, and how we are to explain it…are entirely different functions" (1980, 159). *Ceteris paribus* laws used in scientific theories are literally false, but still do explanatory work. One way to understand Cartwright's claim is that the speech act of describing the world truly and the speech act of explaining come apart from one another. In coming apart from one another and fulfilling different aims within scientific practice, descriptive and explanatory speech acts have different felicity conditions. For example, Potochnik (2016) examines the ways in which explaining increases understanding. But, Potochnik argues, what gets explained depends on a speaker's and audience's interests, and an explaining act's success in generating understanding depends on the cognitive resources of the audience. As such, to evaluate any given communicative act of explaining requires attending to the epistemic and nonepistemic interests of speakers and audiences that form the background against which explanations are offered. This means evaluating explanatory speech acts solely in terms of truth or falsity is inapt.

2.2 Multiple aims and the true/false fetish

I do not think this focus on acts and away from the truth or falsity of descriptive statements is unique to philosophers of science interested in explanation. We see a similar shift in work on the so-called aims approach to values in science (e.g., Elliott and McKaughan 2014;

Intemann 2015). The aims approach shares in common with work on explaining a recognition that scientific practice aims at more than describing the world truly or falsely. Further, if some of those aims include things like making timely policy recommendations for decision makers or increasing public understanding of science, there is a role for nonepistemic values in parts of scientific practice. As Kevin Elliott and Daniel McKaughan put this point, "representations can be evaluated not only on the basis of the relations that they bear to the world but also in connection with the various uses to which they are put" (2014, 3).

Why look to speech act theory to flesh out this picture about the multiple aims of scientific practice and their relationship to nonepistemic values? In part because speech act theory makes sense of the different uses to which one and the same sentence might be put depending on the aims of the speaker and audience and the context of utterance. In doing so, I think Austin is right that we can "play Old Harry with two fetishes…(1) the true/false fetish, (2) the value/fact fetish" (1962, 150). Austin was mainly content to play Old Harry with these fetishes to free philosophers from the grip of the so-called descriptive fallacy: the view "that the sole business, the sole interesting business, of any utterance…is to be true or at least false" (1970, 233). But I also think that in combating the descriptive fallacy and the true/false and fact/value fetishes, speech act theory motivates a constructive shift from the truth or falsity of descriptive statements to the things we do with words.

Take Austin's claim that evaluating apparently descriptive speech acts like "'France is hexagonal,'" involves nonepistemic questions about who is uttering the statement, in what context, and with what "intents and purposes" (1962, 142). Rather than concluding the sentence is false and leaving it at that, Austin points out the different speech acts one can use such a sentence to perform, e.g., stating or interpreting or estimating. In determining the use the sentence is put to—with the help of context and by inquiring after the interests of the speaker and their audience—we might realize, irrespective of the sentence's literal truth or falsity, "It is good enough for a top-ranking general, perhaps, but not for a geographer" (142). In other words, it serves the aims of the general, which, unlike the aims of the geographer, do not necessarily require a descriptively literal account of France's shape. The statement might not aim to assert or describe literally, but do something else entirely. As such,

evaluating it along the lines of truth or falsity will miss something important about the aims of a speaker in uttering it.

To expand on this picture, I turn to explicating Austin's speech act theory.

3. Austin's speech act theory

3.1 Performatives and constatives

Austin first drew our attention to the things we do with words by discussing performative utterances. Austin says of these, "if a person makes an utterance of this sort we should say that he is *doing* something rather than merely *saying* something" (1970, 235). Imagine a speaker utters 'I promise to return my referee report in two weeks' during the peer review process. In making this speech act, Austin claims the speaker does not describe an internal act she has concurrent to her utterance. Instead, in making that utterance, the speaker just is performing the act of promising thereby committing herself to actions related to the timely review of papers.

While promising has no special connection to truth and falsity, it still must meet what Austin calls felicity conditions to be happy or unhappy. In order to promise to return their referee report in two weeks successfully, the speaker must meet the sincerity condition of forming an intention to do so, even if they are not describing "some inward spiritual act of promising" (236). The speaker must also be in a position to follow through on their intention. Thus, there is unhappiness in the speech act if the speaker promises knowing full well other commitments will prevent her from returning the report in two weeks. The speaker must also have the authority to make a promise; unless authorized, an editor cannot promise on behalf of a reviewer. There should also exist a convention for making a promise in peer review contexts. Such conventions might allow the speaker to promise without uttering, 'I promise,' e.g., by accepting a request that reads, 'In accepting this review assignment you commit to returning the referee report within such-and-such a time.'

Austin first contrasts performatives with constatives, e.g., descriptive statements or assertions that aim to state something truly or falsely about the world, but which do not seem to perform an action. However, Austin claims describing or asserting is as much an action as promising, even if the felicity conditions for asserting are more closely connected to truth

or falsity. Consider an editor saying of a reviewer, 'They review quickly, and I expect that they will return their review within two weeks.' In saying this, the editor commits herself to providing evidence for her description of the reviewer as quick, and perhaps justifying her expectation that the reviewer's past behavior provides good evidence for future behavior. As Robert Brandom puts this point, "In asserting a claim one not only authorizes further assertions, but commits oneself to vindicate the original claim, showing that one is entitled to make it" (1983, 641). That is, the utterer must be in a position of authority—here in an epistemic sense—with regards to the claim and be ready to perform further speech acts if so prompted. Other felicity conditions of assertions or descriptions include a sincerity condition: an editor uttering our example sentence should believe what they say. Finally, the context of an assertion also shapes its felicity conditions: an editor should utter the sentence in the appropriate circumstances, e.g., as a response to a worry about the speed of the review process. Should these conditions not be met, the speech act might be unhappy even if true.

3.2 Locution and illocution

Austin develops speech act theory to capture the similarities between performatives and constatives. Speech acts like promising and describing have three dimensions: the locutionary content, which is the conventional sense and reference of the uttered sentence; the illocutionary force, which is the use the utterance is put to; and the perlocutionary effects, which are intended and unintended "effects upon the feelings, thoughts, or actions of the audience, or of the speaker, or of other persons" (1962, 101).

Austin's points about the illocutionary dimension of a speech act most clearly capture how one and the same representation might be put to different uses depending on our goals, and how different uses have different felicity conditions despite sharing locutionary content. Consider the sentence, 'This product contains chemicals known to the state of California to cause cancer.' The locutionary content would just consist in the proposition expressed by the sentence as determined by the conventional sense and reference of the words. This content can be common to different illocutionary acts. Someone uttering the sentence could be describing a product, issuing a warning, or explaining why they do not use this particular product but another. Uttering the sentence with the force of a description, the force of a

warning, and the force of an explanation will have similar felicity conditions related to truth and falsity. Namely, the locutionary content should be true or approximately true for an utterance to count as a good description, a good warning, or a good explanation.

However, a warning might be infelicitous in ways a description might not. For example, warnings might be issued only in the case in which some pre-determined level of significant risk at a certain level of exposure is met. In cases where such levels are not met, issuing a warning might be infelicitous. Consider also that uttering such a sentence with the force of an explanation might be called for only if, e.g., someone is prompted to justify their choice of a product that does not contain cancer-causing chemicals over a more easily available and cheaper product that does contain those chemicals. In these last two cases, nonepistemic reasons related to risk, cost-effectiveness, and so on can enter into the evaluation of the happiness of a warning or explanation.[2]

Austin thinks attending to these points combats a form of abstraction that distorts our thinking about the felicity conditions of descriptive statements. He thinks that when examining statements, "we abstract from the illocutionary…aspects of the speech act, and we concentrate on the locutionary" (1962, 144–5). In so doing, "we use an over-simplified notion of correspondence with the facts—over-simplified because essentially it brings in the illocutionary aspect" (145). Such an approach focuses on "the ideal of what would be right to say in all circumstances, for any purpose, to any audience, &c." (145). But, as Austin claims, questions concerning correspondence with the facts brings with it the illocutionary aspect since truth or falsity does not attach to sentences or locutionary content. Instead, truth or falsity is related to particular things speakers do with sentences. Descriptions might be, strictly speaking, true or false, but not recommendations or explanations. In order to know, then, if evaluating a speech act along the true-false dimension is apt, we need to know the illocutionary force of that act. But to know the illocutionary force of the act requires we attend to context, including the aims of both speaker and audience, time and place of utterance, and conventions governing the specific speech situation. In this way, Austin

[2] Any speech act will also have perlocutionary effects, and we might follow Heather Douglas (2009) and Paul Franco (2017) in focusing on the nonepistemic consequences of making false descriptions, giving bad warnings, or explaining unclearly.

argues context and aims are central to determining the illocutionary force of a speech act, and hence to evaluating its felicity or infelicity.

4. Aims-approaches and speech act theory

4.1 Explaining and understanding

Scientific practice might seem to deal in paradigmatically constative speech acts, e.g., descriptions. Such speech acts are, to varying degrees, evaluable along dimensions of truth or falsity in ways we might question the relevance of speech act theory to philosophy of science. That is, we might say that scientific practice just is a case in which abstracting away from the illocutionary force of an utterance to focus on locutionary content is appropriate. For example, Austin says that "perhaps with mathematical formulas in physics books…we approximate in real life to finding" speech acts where focusing on the locutionary content is appropriate (1962, 145). If scientific practice aims at timeless truths holding across all contexts independent of the sorts of aims and interests of speakers and audiences necessary to evaluating the felicity or infelicity of speech acts, then it seems speech act theory is irrelevant to philosophy of science.

Yet, as Austin points out, "When a constative is confronted with facts, we in fact appraise it in ways involving the employment of a vast array of terms which overlap with those that we use in the appraisal of performatives. In real life, as opposed to the simple situations envisaged in logical theory, one cannot always answer in a simple manner whether it is true or false" (141–2). Consider again 'France is hexagonal.' Austin asks, "How can one answer…whether it is true or false that France is hexagonal? It is just rough, and that is the right and final answer to the question of the relation of 'France is hexagonal' to France. It is a rough description; it is not a true or false one" (142). Though rough, it is still open to evaluation. We can ask if it is in accord with conventions governing estimations and if this estimation serves the purposes and interests of the speaker and their audience at the time of utterance. 'France is hexagonal' can count as felicitous even if rough and not literally true because it aims at something other than truth.

Austin claims that many of our apparently constative speech acts are evaluable along similar dimensions given that they also confront facts in similarly rough ways. McKaughan

makes a related point about scientific speech acts. He argues that certain speech acts central to scientific practice like "conjecturing, hypothesizing, guessing and the like often play a role in scientific discourse that serves neither to assert that an hypothesis is true nor to express such a belief" (2012, 89). Moreover, as mentioned in §2, the picture of scientific practice as concerned solely with the truth is challenged, among other places, in work on explanation, and also in values in science. For example, when looking at the role particular acts or patterns of explaining play in scientific discourse we might focus not on the locutionary content of an explanatory speech act, but on the ways "explanatory discourse…functions to sculpt and subsequently perpetuate communal norms of intelligibility" (Woody 2015, 81). In focusing on this aspect of explaining, we might find, for example, that "the ideal gas law's role in practice is not essentially descriptive, but rather prescriptive; by providing selective attention to, and simplified treatment of, certain gas properties (and their relations) and ignoring other aspects of actual gas phenomena, the ideal gas law effectively instructs chemists in how to think about gases as they are characterized within chemistry" (82). In other words, the ideal gas law, in practice, does not have the force of a descriptive speech act, but lays down a rule of sorts guiding the investigation of gases.[3] The success of acts of explaining from this perspective will have less to do with accurately describing actual gases, but the way they facilitate, say, the education of new scientists or increase understanding of related phenomena, e.g., "by laying foundation for the concept of 'temperature'" beyond "the subjective, inherently comparative quality of human perception" (82). An act of explaining that fails to achieve pedagogical aims or fails to increase understanding of related phenomena might be infelicitous even if the locutionary content of that act confronts the facts in the right way to count as approximately true.

On this point about the ways explanations might increase understanding without describing, Potochnik claims "that what best facilitates understanding is not determined solely by the relationship between a representation and the world" (2015, 74). An idealized explanation like the ideal gas law is not defective because it fails to fully describe all the

[3] About universal generalizations Austin writes, "many have claimed, with much justice, that utterances such as those beginning 'All…' are prescriptive definitions or advice to adopt a rule" (1962, 143). Austin does not fully endorse this suggestion.

possible causal factors at play in the behavior of actual gases. Though literally false, an idealization might be successful insofar as it "secure[s] computational tractability" or successfully isolates "all but the most significant causal influences on a phenomenon" (71). In so doing, we increase our understanding by facilitating "successful mastery, in some sense, of the target of understanding" or "by revealing patterns and enabling insights that would otherwise be inaccessible" (72). Indeed, pointing out all the ways in which the ideal gas law fails to hold for actual gases or is literally false as a description might hinder the use of explanations in scientific discourse to provide "shared exemplars that function as norms of intelligibility" (Woody 2015, 84).

In a related vein, Potochnik argues, "Because understanding is a cognitive state, its achievement depends in part on the characteristics of those who seek to understand," including both the speaker and the audience (2015, 74). In evaluating an act of explaining, we should look at how the speaker's interest has shaped the focus of their explanation and also how the explanation increases an audience's understanding, where this involves considering the audience's interests in seeking an explanation. An explanation that fails to be relevant to the audience or fails to increase their understanding or guide their thinking about related phenomena, but that nonetheless has locutionary content that is approximately true, might count as infelicitous.

4.2 Values and science

On the views of explaining canvassed, the aims of generating literally true descriptions of the world come apart from, say, explaining and understanding the most important causal factors at play for a given phenomenon. Now, as the aims approach to the proper role for nonepistemic values in scientific practice emphasizes, explaining and describing do not exhaust the goals of scientific practice. The aims approach focuses on the ways "scientific decision-making, including methodological choices, selection of data, and choice of theories or models, are...a function of the aims that constitute the research context" (Intemann 2015, 218). Given that the research context includes social, political, and moral considerations, the aims of science can just as well be understood in nonepistemic ways as it can be understood in epistemic ways.

Consider, for example, the American Geophysical Union's position statement on human-induced climate change. At the end of their statement, they claim, "The community of scientists has responsibilities to improve overall understanding of climate change and its impacts. Improvements will come from pursuing the research needed to understand climate change, working with stakeholders to identify relevant information, and conveying understanding clearly and accurately, both to decision makers and to the general public" (American Geophysical Union 2013). Here, I focus on the claim that scientists have responsibilities to improve the understanding of policymakers and the general public, and drawing upon the aforementioned work on explaining, think about how adopting this aim shapes the felicity conditions of explanatory speech acts directed at the audiences mentioned.

Notice that the position statement distinguishes the research necessary to understand climate change from conveying that understanding to policymakers and the general public. The sense in which these different activities come apart from one another and have different success conditions can be made sense of, in part, by focusing on the audience to whom scientists are speaking. We saw that for Potochnik (2016) understanding is a cognitive state that depends on the abilities and interests of those who are explaining and those to whom explanations are directed. In communicating to policymakers and the general public, scientists should consider the interests of the speaker in asking for an explanation as well as their level of knowledge regarding the phenomenon in question, in this case, climate change. In so doing, scientists might find that a description that aims to describe climate change in all its complexity might not serve these aims well. Instead, scientists might aim for an explanation that, though omitting descriptive complexity, draws upon models that represent those causal factors related to the audience's interests in a way that is cognitively accessible and helps guide the public in thinking more generally about climate change.

On this point, the American Geophysical Union's position statement maintains scientists ought to enlist the help of stakeholders in identifying potentially relevant information to their research. This is a point Intemann makes in developing the aims approach. She says of climate science, "[T]he aim is not only to produce accurate beliefs about the atmosphere, but to do so in a way that allows us to generate useful predictions for protecting a variety of social, economic and environmental goods that we care about" (2015,

219). In the view of the American Geophysical Union, in order to do this well, scientists ought to consult with relevant stakeholders and policymakers regarding what they value. Thus, for example, if stakeholders and policymakers communicate worries about extreme weather events and "how to adapt to 'worst case scenarios,' then models able to capture extreme weather events should be preferred" to those models that "anticipate slow gradual changes" (Intemann 2015, 220). Notice that in making such a decision, the grounds for choosing models able to represent aspects of climate change relevant to stakeholders' interests are nonepistemic rather than epistemic, e.g., generating predictions useful for protecting goods the general public cares about. Insofar as the representations or explanations generated do not meet these goals because they are unrelated to stakeholders' interests, the attendant speech acts might very well be infelicitous even if they describe some related phenomenon more or less accurately.

Both points about pitching explanations at a level that is cognitively accessible and choosing models for representing climate change phenomena in ways sensitive to stakeholders' interests illustrate a point Austin makes about the importance of uptake to successfully performing a speech act. Austin claims, "Unless a certain effect is achieved, the illocutionary act will not have been happily, successfully performed….I cannot be said to have warned an audience unless it hears what I say and takes what I say in a certain sense….Generally the effect amounts to bringing about the understanding of the meaning and force of the locution" (1962, 116). In aiming to convey understanding through explaining relevant aspects of climate change to decision makers and the general public, a speaker should consider the interests, background knowledge, and cognitive resources of their audience. Insofar as scientists fail to do so in explaining to the general public, even if the locutionary content that comprises their speech act approximates truth, they will not secure uptake in the sense of generating understanding in their audience. As such, their speech act will be infelicitous.

Of course, a scientist's explaining something to their audience will also be infelicitous if it is based on inaccurate information or extrapolates from what is known to their audience's interests in unjustified ways. However, this does not mean that if scientists aim to convey understanding to the public they should stick solely to descriptive claims. As

Elliott emphasizes in discussing how scientists should best communicate uncertainty to the public, "It does little good to expect scientists to provide unbiased information to the public if their pronouncements are completely misinterpreted or misused by those who receive them" (2017, 89). Similarly, "members of the public might not be able to 'connect the dots'" between scientists' descriptive speech acts and the ways those are relevant to their interests; insofar as scientists do not explain with the aims of conveying understanding—which as Potochnik argues, comes apart from describing the world truly in all its complexity—the public "would be left wondering what [the descriptions] might mean" (88). Thus, if scientists are to meet responsibilities the American Geophysical Union claims they have with regard to conveying understanding to the general public, those scientists should communicate using speech acts best able to secure uptake in the general public. This involves considering the interests and cognitive resources of the general public in ways that shape the felicity conditions of the speech acts beyond truth and falsity.

5. Conclusion

I argued speech act theory can tie together a few threads in recent work on explaining and values in science that share in common a shift in focus from descriptive propositions to things scientists do with words. Some of those things, like explaining, also seem the sorts of speech acts appropriate for fulfilling aims scientists have other than describing the world literally, like conveying understanding to the public and policymakers. Insofar as successfully fulfilling these aims involves explaining, and insofar as acts of explaining that secure uptake require attention to the nonepistemic interests and cognitive resources of speaker and audience, our attention is drawn towards ways explanatory speech acts can be happy or unhappy beyond describing truly or falsely. Future work will aim to delineate these felicity conditions in greater detail with an eye towards revealing further nonepistemic dimensions of evaluation.

References

Achinstein, Peter. 1977. "What is an Explanation?" *American Philosophical Quarterly* 14(1):1–15.

American Geophysical Union. 2013. "Human-Induced Climate Change Requires Urgent Action." https://sciencepolicy.agu.org/files/2013/07/AGU-Climate-Change-Position-Statement_August-2013.pdf

Austin, J.L. 1962. *How to Do Things With Words*. Ed. J.O. Urmson. Oxford: Oxford University Press.

----. 1970. "Performative Utterances." *Philosophical Papers*, 2nd edition. Eds. J.O. Urmson and G.J. Warnock. Oxford: Oxford University Press: 233–252.

Brandom, Robert. 1983. "Asserting", *Nous* 17(4):637–650.

Cartwright, Nancy. 1980. "The Truth Doesn't Explain Much." *American Philosophical Quarterly* 17(2):159–163.

Chang, Hasok. 2014. "Epistemic Activities and Systems of Practice: Units of Analysis in Philosophy of Science After the Practice Turn." *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science*, eds. Léna Soler, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost. New York: Routledge: 67–79.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, Kevin. 2017. *A Tapestry of Values*. New York: Oxford University Press.

Elliott, Kevin C. and Daniel J. McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81(1):1–21

Franco, Paul L. 2017. "Assertion, Nonepistemic Values, and Scientific Practice." *Philosophy of Science* 84(1):160–180.

Intemann, Kristen. "Distinguishing Between Legitimate and Illegitimate Values in Climate Modeling." *European Journal of the Philosophy of Science* 5:217–232.

McKaughan, Daniel J. 2012. "Speech acts, attitudes, and scientific practice: Can Searle handle 'Assuming for the sake of Hypothesis'?" *Pragmatics and Cognition* 20:1:88–106.

Potochnik, Angela. 2015. "The Diverse Aims of Science." *Studies in History and Philosophy of Science Part A* 53:71–80

----. 2016. "Scientific Explanation: Putting Communication First." *Philosophy of Science*, 83:721–732.

Woody, Andrea. 2015. "Re-orienting discussions of scientific explanation: A functional perspective." *Studies in History and Philosophy of Science Part A* 52:79–87.

Universality Reduced

Alexander Franklin*†

October 2018

Forthcoming in *Philosophy of Science: Proceedings of the PSA 2018*

Abstract

The universality of critical phenomena is best explained by appeal to the Renormalisation Group (RG). Batterman and Morrison, among others, have claimed that this explanation is irreducible. I argue that the RG account is reducible, but that the higher-level explanation ought not to be eliminated. I demonstrate that the key assumption on which the explanation relies – the scale invariance of critical systems – can be explained in lower-level terms; however, we should not replace the RG explanation with a bottom-up account, rather we should acknowledge that the explanation appeals to dependencies which may be traced down to lower levels.

1 Introduction

While universality is best explained with reference to the Renormalisation Group (RG), that explanation is nonetheless reducible. The argument in defence of this claim is of philosophical interest for two reasons: first, the RG explanation of universality has been touted by Batterman (2000, 2017) and

*alexander.a.franklin@kcl.ac.uk

†I am grateful to Eleanor Knox, and to the audience of the IMPS 2018 conference in Salzburg for helpful comments. This work was supported by the London Arts and Humanities Partnership.

Morrison (2012, 2014) as a significant impediment to reduction. Second, universality is a paradigm instance of multiple realisability (MR) in the philosophy of physics; as such it is regarded as irreducible by those who accept the multiple realisability argument against reduction. My account charts a middle course: I deny claims that RG explanations are irreducible, and I deny that universality is *best* explained from the bottom up.

The view of reduction advocated here is non-eliminativist; the best explanations are often higher-level explanations: such explanations are more parsimonious, more robust, and have broader applicability than lower-level explanations. In general, such higher-level explanations ought not to be replaced by lower-level explanations, rather the parts of theories on which such explanations rely may be understood in lower-level terms; reducible explanations satisfy the following two conditions: (a) each higher-level explanatory dependency is explained by or derived from a lower-level dependency, and (b) the abstractions involved in constructing the higher-level explanations are justified from the bottom up.[1]

In §2 I outline the RG explanation of universality. Although my reductive claims may generalise, I focus exclusively on the field-theoretic approach to the RG.[2] I claim that this explanation follows a general formula for explaining multiply realised phenomena. §3 considers the arguments of Batterman and Morrison, and analyses their force against any putative reduction.

In §4 I note that the RG explanation is a higher-level explanation. As it is less contentious that the common features of each universality class are reducible, I simply assume that that's the case in this paper. The nub of the debate rests on the RG: I show that the RG arguments rely on the assumption of scale invariance and the abstractions engendered by that assumption. I argue that the applicability of this assumption may be explained from the bottom up. Thus, I claim, that my reduction satisfies (a) and (b) above.

[1]While I expect the claims in this paper to be compatible with many different accounts of explanation, they are most straightforwardly cashed out on an interventionist approach – see Woodward (2003).

[2]See Franklin (2018) and Mainwood (2006) for arguments that only this approach provides an adequate explanation of universality.

2 The RG Explanation of Universality

'Universality' refers to the phenomenon whereby diverse systems exhibit similar scaling behaviour on the approach to a continuous phase transition. Continuous phase transitions occur at the critical temperature, a point beyond which systems no longer undergo first-order phase transitions.[3] The approach to this phase transition can be very well described by power laws of the form $a_i(t) \propto t^{\alpha}$ where t is proportional to the temperature deviation from the critical temperature and α is the critical exponent – a fixed number which leads to a characteristic curve on temperature-density plots.[4]

Different physical systems can be categorised into universality classes: members of the same class have identical critical behaviour – the same set of critical exponents $\{\alpha, \beta, ...\}$ for several power laws – while their behaviour away from the critical point and microscopic organisation may be radically different. For example, fluids and magnets are in the same universality class despite otherwise having totally different chemical and physical properties.

Each physical system which exhibits critical phenomena may be described at the critical point by the same mathematical object – the Landau-Ginzburg-Wilson (LGW) Hamiltonian. That Hamiltonian will include the features – the symmetry and dimensionality – which sort these systems into their universality classes. The RG argument demonstrates that the LGW Hamiltonian applies to a wide range of systems at the critical point by showing that any additional operators which may be appended to that Hamiltonian will fall away on approach to criticality, where only the central LGW operators will remain. The following steps are essential to the explanation thus on offer:[5]

1. Define the effective Hamiltonian for your system of interest:
   (i) Specify the order parameter with symmetry and dimensionality.
   (ii) Specify the central operators of the LGW Hamiltonian.

[3] Note that not all continuous phase transitions are associated with first-order phase transitions in this way.

[4] E.g. the specific heat (in zero magnetic field) c scales as $c \sim (t^{-\alpha})/\alpha$ as $t \to 0$ where $t = \frac{T - T_c}{T_c}$.

[5] To see a full account of the physics of universality and details of the RG see Binney et al. (1992) and Fisher (1998); the philosophical aspects of such an explanation are discussed in detail in Batterman (2016) and Franklin (2018).

   (iii) Specify operators in addition to the terms in the LGW Hamiltonian.

2. Apply the RG transformations to that Hamiltonian.

3. Examine the flow towards fixed points in the critical region and note that some operators are irrelevant to the critical behaviour.

4. Thus divide the set of operators into subsets: 'relevant', 'irrelevant' and 'marginally relevant'.

5. Repeat for other systems of interest.

In order to explain universality we must identify commonalities between the different systems in the same universality class – 1(i) and 1(ii) above – and show that such commonalities are sufficient for the common behaviour – 2-4 above. Although 1(iii) can't, in general, be done explicitly, the explanation only depends on the RG demonstration that all distinguishing features are irrelevant – it's not necessary to say exactly which those distinguishing features are. As discussed below, the infinities which are central to some of the anti-reductionist arguments feature in steps 3 and 4.

Overall the explanation takes the following form: consider a universality class composed of four different physical systems A-D. Each of A-D is described in step 1 by an effective Hamiltonian; effective Hamiltonians are ascribed to systems on the basis of various theoretical and empirical data. The RG explanation of universality, by virtue of steps 2-4, tells us that all the details which distinguish A-D, i.e. their irrelevant operators, are, in fact, irrelevant to the critical phenomena. Thus we have an explanation for how otherwise different systems exhibit the same phenomena at the critical point. This explanation relies, of course, on the RG transformations which allow for the categorisation of certain operators as irrelevant.

Importantly, this explanation takes the form of a general explanation of multiply realised phenomena: such phenomena are explained if commonalities are identified among the realisers and these are shown to be sufficient for the multiply realised phenomena to occur. Note that such explanations may be higher level and nothing written so far establishes their reducibility.

3 Anti-reductionist Arguments

Batterman (2000, 2017) and Morrison (2012, 2014) offer two arguments in defence of the view that the explanation just outlined is irreducible. The more general argument is that universality, *qua* instance of multiple realisability, is irreducible because multiple realisability requires abstracted explanations of a particular form.

However, one goal of this paper is to demonstrate that just such abstracted explanations may be reducible. Insofar as my reduction of the RG explanation goes through, we are thus faced with a dilemma: either some instances of MR are, in principle, reducible, or universality is not a case of MR. While I would opt for the former horn, nothing in the rest of the paper hangs on that choice.

The second anti-reductionist argument is much more specific to the case at hand and involves various demonstrations that the RG explanation requires infinities which are inexplicable from the bottom up. As noted by Palacios (2017), two different limits are invoked in the case of continuous phase transitions – the thermodynamic limit and the limit of scale invariance. There is an extensive literature on the thermodynamic limit as it appears in first order phase transitions; as I see no salient differences between appeal to this limit in the two contexts, I do not discuss this further here – see e.g. Butterfield and Bouatta (2012) for a reductionist account of that limit.[6]

The second limit is discussed by Butterfield and Bouatta (2012), Callender and Menon (2013), Palacios (2017), and Saatsi and Reutlinger (2018), among others, and these papers undermine claims that continuous phase transitions are irreducible. However, they pay insufficient attention to the specific role played by the RG (and by the limit of scale invariance) in establishing the irrelevance of certain details, and it is this role which is crucial to the anti-reductionist arguments.[7]

For Batterman, the RG is required because it allows us to answer the following question:

[6]The reductionist claims made here are conditional on a successful resolution of such issues.

[7]For example, Saatsi and Reutlinger (2018, p. 473) do not consider a counterfactual of the form 'if a physical system S did not exhibit effective scale invariance at criticality, then S would not exhibit the critical phenomena of any universality class' in their list of counterfactuals which the RG account is supposed to underwrite.

> **MR**: How can systems that are heterogeneous at some (typically) micro-scale exhibit the same pattern of behavior at the macro-scale? ...
>
> if one thinks (**MR**) is a legitimate scientific question, one needs to consider different explanatory strategies. The renormalization group and the theory of homogenization are just such strategies. They are inherently multi-scale. They are not bottom-up derivational explanations.
>
> [Batterman (2017, pp. 4, 14-15)]

As further elaborated below, the RG seems to Batterman to preclude "bottom-up derivational explanation" because it requires the following infinitary assumption:

> This [fixed point] is a point in the parameter space which, under τ [the RG transformation], is its own trajectory. That is, it represents a state of a system which is invariant under the renormalization group transformation. Of necessity, such a fixed point has an *infinite correlation length* and so lies on the critical surface S_∞. The singularity/divergence of the correlation length ξ is *necessary*.
>
> [Batterman (2011, p. 1045), original emphasis]

I accept that the RG formalism makes use of infinite limits. The salient question, to borrow Norton's (2012) distinction, is whether such infinities are approximations which allow one to use the more tractable infinitary mathematics to approximate features of the finite systems, or, alternatively, idealisations which describe a distinct infinite system. Claiming that the infinities are idealisations would preclude reduction because the macroscopic system with infinite properties has features which may not be reductively explained.

As Batterman demonstrates, the RG argument rests on the assumption of the infinite correlation length which generates absolute scale invariance. In §4 I claim that the physical systems under consideration are not absolutely scale invariant: in fact, one may abstract from the details of the underlying system insofar as such systems are effectively scale invariant; thus the infinitary assumption is best viewed as an approximation.

While Morrison (2014, p. 1155) likewise focusses on explanations of MR phenomena, she claims that RG explanations are irreducible for a different, but related, reason: the "RG functions not only as a calculational tool but as the source of physical information as well". Morrison (2012) makes a similar argument in relation to symmetry breaking in the physics of superconductors. She argues that, in both cases, top-down constraints play an essential role in the physical descriptions which thus rules out reduction. In the present context, Morrison's views may be understood as taking the RG invocation of scale symmetry to be a necessary physical assumption which cannot be understood from the bottom up. Below I argue that the effective scale invariance on which the RG rests is, in fact, reductively explicable. As such, no top-down organising principles are required and Morrison's claims are deflated.

4 Reducing the RG Explanation

Arguments for the reducibility of the explanation of universality have primarily been targeted at Batterman's claims that infinities are essential to the models used to describe continuous phase transitions. I do not have space to consider these arguments in any detail. Suffice it to say that, in my view, none succeeds in reducing the principal feature of the renormalisation group – the assumption of scale invariance. Thus I focus on that aspect of the RG, and claim that it, too, is reducible.

Furthermore, with the notable exception of Saatsi and Reutlinger (2018), not much attention has been paid to the explanation of universality *per se*. This, of course, makes a difference for MR-based objections to reduction, which raise doubts that a reductionist account could explain why the same phenomenon is exhibited in multiple different systems.

As far as the physics is currently developed, the RG plays an ineliminable role in the explanation of universality: it is the only mathematical framework available to predict the precise extent of observed universality of critical phenomena. If its application were truly mysterious, if we had no idea why it worked, then, infinity or no infinity, this would provide exactly the right kind of failure of explanation on which the anti-reductionist could hang their arguments.

I argue in the following that the applicability of the RG to systems un-

dergoing continuous phase transitions is not mysterious. The RG exploits effective scale invariance to set up equations which tell us how certain properties vary with respect to the variation of other properties. It is a piece of mathematics whose applicability is deeply physical – where the assumptions invoked in applying the RG do not hold, the RG's predictions go wrong.

In order fully to reduce the RG explanation, one also must consider the common features shared by each member of the same universality class, and argue that these, too, are reducible to aspects of the microphysical description. Such arguments have been given by the reductionists mentioned above. The innovation of this paper lies in reducing the RG framework, and the assumptions on which it relies; thus, given space constraints, I do not consider the reduction of the symmetry, dimensionality and representation by common Hamiltonians.

4.1 Reducing the Renormalisation Group

The RG argument rests on the assumption of scale invariance, and this is crucial to the demonstration that a class of operators are irrelevant at criticality. I claim that we can provide a bottom-up explanation of this scale invariance and that, as such, the RG arguments provide a mathematical apparatus for relating scale invariance to the irrelevance of certain details. One can see, heuristically, how scale invariance relates to universality: if the system at criticality is effectively scale invariant then many of that systems' features – those which are scale dependent – will turn out to be irrelevant at criticality, and all that will remain are those shared features such as the symmetry and dimensionality.

To argue that the RG explanation is reducible, I first give a more general characterisation of an RG flow. The calculation of each system's dynamics involves integration over a range of scales and energies. The highest energy (smallest scale) cutoff (denoted Λ) corresponds to the impossibility of fluctuations on a scale smaller than the distance between the particles in the physical system. The RG transformation involves decreasing the cutoff thereby increasing the minimum scale of fluctuations considered. Iterating this transformation generates a flow through parameter space designed to maintain the Hamiltonian form and qualitative properties of the system in question.

The RG transformation $\mathcal{R}$ transforms a set of (coupling) parameters $\{K\}$ to another set $\{K'\}$ such that $\mathcal{R}\{K\} = \{K'\}$. $\{K^*\}$ is the set of parameters which corresponds to a fixed point, defined such that $\mathcal{R}\{K^*\} = \{K^*\}$. This fixed point corresponds to the critical point defined physically. At the fixed point, the RG transformation (which changes the scale of fluctuations) makes no difference. Thus the fixed point encodes the property of scale invariance.

Given the Hamiltonian of one of our models, one can define an RG transformation which generates a flow that allows one to: (i) classify certain of the coupling parameters of the system in question as (ir)relevant to its behaviour near the fixed point, (ii) extract the critical exponents from the scaling behaviour near the fixed point.

The RG may be understood as a mathematical framework for exploring how certain properties vary with changing energy, length-scale, or, by proxy, temperature, on approach to the scale invariant critical point. Philosophical discussions of the RG are occasionally prone to mysterianism, but the RG should be considered to be no different from, for example, the calculus. As Wilson (1975, p. 674) notes: "the renormalization group ... is the tool that one uses to study the statistical continuum limit [the point of scale invariance] in the same way that the derivative is the basic procedure for studying the ordinary continuum limit".

The Hamiltonian which represents the system at the critical point, from which the critical exponents are extracted, is scale invariant at the fixed point – all the scale dependent contributions have gone to zero. Such Hamiltonians are known as 'renormalisable'. As such, the explanation provided below for the effective scale invariance of physical systems at criticality underlies the fact that such systems are well-described by renormalisable Hamiltonians at fixed points.

My argument has two steps: I demonstrate that scale invariance is implicit in the power law behaviour which is intrinsic to universality; then I provide a bottom-up explanation of the effective scale invariance for liquid-gas systems, a story somewhat motivated by the observation of critical opalescence. Thus, I show how scale invariance features in the mathematics – the Hamiltonian's renormalisability and the power laws, and how it features in the observed physics – the critical opalescence is a direct consequence of the bottom-up story.

The universality of critical phenomena lies in the sharing of power laws,

and hence critical exponents, between members of the same universality class. In what sense are such power laws scale-free? As Binney et al. (1992, p. 20) explain, a phenomenon obeying a power law is independent of scale because one could multiply its characteristic scale length by some factor and the ratio of values will remain constant. For example, consider the power law $f_1 = (r/r_0)^\eta$, and its measurement in the range $(0.5r_0, 2r_0)$. The ratio of largest to smallest value will be identical for measurements centred on $r_0, 10r_0, 100r_0$ – it will always be $4^{|\eta|}$, thus one may superimpose all the power laws by a simple change of scale. By contrast, for $f_2 = \exp(r/r_0)$ the ratio of values will change on scale changes.

Such systems are therefore described as scale-free; the RG is used to predict that at the point of scale invariance the heterogeneous features will be irrelevant. So, in order to work out when this framework is applicable, and why it works, we ought to look at each individual system, (for our purposes let's reserve inquiry to liquid-gas and ferromagnetic-paramagnetic systems) and identify the underlying processes which lead to effective scale invariance at the critical point. The following two caveats apply to this proposal for reduction:

First, it might be objected that universality may only be explained if the same processes are identified across all the systems exhibiting the universal behaviour; if that were so, the strategy employed here would be inadequate. However, universality may be explained by demonstrating that two conditions are fulfilled: that all the systems share common features, and that their heterogeneous details are irrelevant. While it's essential that the common features are shared by all the systems, the mechanism by which the heterogeneities are irrelevant may differ, so long as all the heterogeneities in fact end up as irrelevant.

Second, although the power laws and renormalisable Hamiltonians at the fixed point are absolutely scale invariant, the physical systems will, at best, be effectively scale invariant – that is, scale invariant within a certain range of length-scales. That should be acceptable because we know that scale invariance is never exactly true of a system: any real system will be finite and thus violate the assumption at some scale. Moreover, this will not generate empirical problems because the power laws are observed for systems approaching criticality – they are predictions about $T \to T_c$, not $T = T_c$. Thus one should only assume that critical exponents asymptotically approach those predicted at the fixed point. While infinite assumptions are required in order to impose the full scale invariance for RG analy-

sis, I claim that we can explain effective scale invariance for finite systems, and that absolute scale invariance is an approximation invoked to make the mathematics tractable.

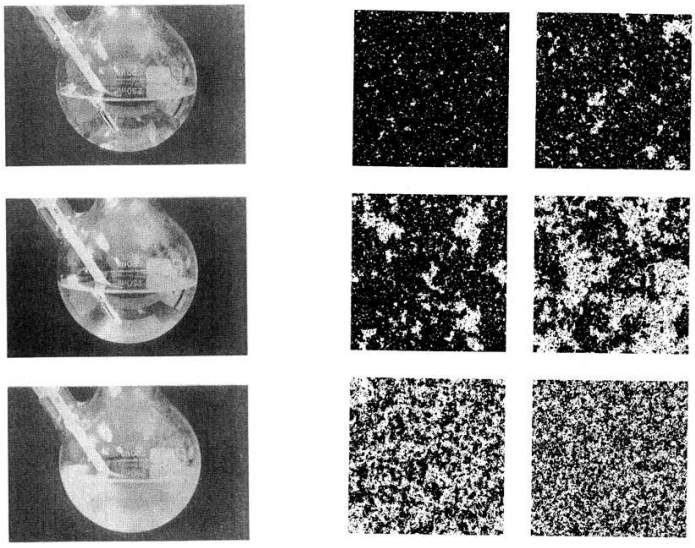
Scale invariance, as it manifests in systems at criticality, is known as 'self-similarity': as scales change the system resembles itself. How do we account for such self-similarity? The critical point, at which a continuous phase transition occurs, corresponds (for liquid-gas systems) to the highest temperature and pressure at which liquid and gas phases can be distinguished.

As is well known, there is a plateau in pressure-volume diagrams, which corresponds to the latent heat (or enthalpy) of vapourisation. This, roughly, is the extra energy needed to break the intermolecular bonds which distinguish liquids from gases and vapours. At the critical point this plateau, and the latent heat of vapourisation vanishes. Now it's difficult precisely to work out the binding energies of the intermolecular bonds. The values for this will be material dependent, and surface tension dependent, and will change at different pressures. But the heuristic argument tells us that the reason the plateau vanishes is because the system has enough temperature, and thus the molecules have sufficient energy to equal the binding energy. The point at which binding energy is exactly matched by kinetic energy will be the critical point.

The isothermal compressibility (κ) is defined as $\kappa_T = \frac{-1}{V}\left(\frac{\partial V}{\partial p}\right)_T$. This corresponds to how much the volume will change (∂V) with a given pressure change (∂p) at fixed temperature (T). As supercritical fluids have far higher compressibility than liquids, and both are present at the critical point, the compressibility diverges. Given, in addition, that the latent heat is zero at criticality, there's nothing to prevent a given bubble expanding arbitrarily. Thus we ought to expect the system to have bubbles of all sizes: this is what is meant by the claim that the system is dominated by fluctuations and has no characteristic scale.[8]

Negligible energy cost for transitions and infinite compressibility leads to self-similarity, and, in certain fluids, the bubbles at all scales lead to a high refraction of visible light. Thus otherwise transparent fluid may become opaque and milky-white. This is known as 'critical opalescence' – see figure 1(a) – and is a visible correlate of a system at criticality.

[8]Note that, for first order phase transitions, the compressibility also diverges; this doesn't lead to scale invariance because latent heat is finite.



(a) (b)

Figure 1: From Binney et al. (1992, pp. 10,19). (a) Critical opalescence is visible when arbitrarily large bubbles form in liquid at criticality. (b) Increasing loss of characteristic scale as $T \to T_c$ in simulations of the Ising model.

Such self-similarity is conceptually crucial to the applicability of the renormalisation group: in order to extract critical exponents from RG equations one identifies a renormalisable Hamiltonian which is scale invariant at the fixed point. Without fluctuations across all scales, systems would fail to be well modelled by such Hamiltonians. The physical argument for diverging fluctuation size justifies the use of a scale invariant mathematical model to represent such systems. Thus, for critical phenomena, the applicability of the RG depends on scale invariance, where this assumption is explicable from the bottom up.

Demonstrating these claims quantitatively is difficult, but the heuristic argument is convincing. Kathmann (2006) reviews theories of the nucleation of gas bubbles in water which generate accurate predictions concerning the rate of bubble growth and the threshold for stability over a range of temperatures; although these models do not reach the critical point, progress is being made.[9]

[9]Constructing exact models is especially difficult because of the fluctuations at a wide range of length scales – precisely the reason that the RG is employed.

Of course, further work could be done to develop these arguments and make them more precise. But there seems to be, in the above, a sound qualitative argument and no in-principle barriers to full derivation. This 'in-principle' ought not to be problematic: we know the relevant physical principles, even if quantitative models are still unavailable.

Moreover, as discussed below, and depicted in figure 1(b), the Ising model allows us quantitatively to predict analogues of the results for liquid-gas systems. While well short of a full explanation, the following discussion illustrates how self-similarity may be reduced for magnetic systems. By treating the Ising model as a stand-in for such systems, a similar kind of reasoning to that given above will go through.

Below the critical point, energy fluctuations will lead to random isolated spin flips. Such flips will be energetically costly and tend to be reversed. The higher the energy, the more likely these are to occur, and if sufficiently many occur then a patch will form, and other spins will have some tendency to align themselves with this patch. However, below the critical point, such patches beyond a certain size will be too costly and spins will overall remain aligned (there is some small probability of net magnetisation flipping, but this is increasingly unlikely further below the critical point).

At the critical point, the energy of the atoms in the lattice is greater than the energetic cost of violating spin alignment, and patches can become arbitrarily large. This results from the latent heat's vanishing and the divergence of the magnetic susceptibility (χ) on approach to the critical point. $\chi_T = \left(\frac{\partial m}{\partial B}\right)_T$ where m is the magnetisation and B represents an external magnetic field. Universality is manifested by the fact that the susceptibility and the compressibility both diverge according to identical power laws with the same critical exponent γ: χ_T, $\kappa_T \sim (T - T_c)^{-\gamma}$. Thus, we have self-similarity and effective scale invariance with bubbles or patches arbitrarily large up to the size of the system.

My aim is to establish the reducibility of the RG relevance and irrelevance arguments. I have demonstrated that the RG is a mathematical procedure that extracts information based on the empirically and theoretically justified assumption of effective scale invariance; this has been shown to be a property shared by different systems at criticality. The key ingredients for effective scale invariance are features of the interactions of neighbouring sub-systems, and the particulate constitution of the materials. While that suggests that these materials are not so different after all, it's worth empha-

sising that the systems which exhibit universal behaviour are nonetheless dissimilar away from the critical point – it's clear that magnets and liquids have many distinct chemical and physical properties.

The assumption of scale invariance plays a crucial role for the RG – it licences the discarding of scale dependent details; it is precisely this discarding of details which ensures that all systems are commonly described at the critical point. Moreover, discarding such details is what gives the higher-level explanation its stability and parsimony. It is thus incumbent on the reductionist to explain how the higher-level RG account is successful despite its leaving out such details. So, the reductionist should identify physical processes at the lower level which ensure the irrelevance of the discarded details.

As argued above, the physical processes in question are exactly those which lead to effective scale invariance. The fluctuations at all scales make it such that the scale-dependent properties which distinguish systems away from criticality are irrelevant at criticality, when the system is effectively scale invariant. We have identified, at the molecular level, the physical mechanisms which prevent variations in the discarded details from leading to changes in the higher-level description of the system. As such, we are assured that the explanatory value of the higher-level explanation is a consequence of features of the lower-level system.

One upshot of this reductionist account is that we may specify the conditions under which the higher-level description remains a good one. The discarded details are irrelevant while the large scale fluctuations – the bubbles or patches – dominate the physics. As we move to systems which are less scale invariant, as the bubbles die down, the critical point becomes a less accurate description and each system in the class will start to exhibit distinct behaviour. This is reflected in the fact that the macroscale RG description only derives the shared behaviour at the fixed point of scale invariance and predicts distinct behaviour away from the fixed point.

I end this section with the following intuitive physical gloss on the RG explanation: "[b]ecause the fluctuations extend over regions containing very many particles, the details of the particle interactions are irrelevant, and a great deal of similarity is found in the critical behavior of diverse systems" (A. L. Sengers, Hocken, and J. V. Sengers (1977, p.42)). Since we can explain the wide-ranging fluctuations from the bottom-up, the RG explanation of universality is reducible.

5 Conclusion

The field-theoretic RG framework, together with the common features of physical systems in the same universality class, explains how those systems all display the same critical phenomena when undergoing continuous phase transitions. That explanation is a higher-level explanation.

That higher-level RG explanation is nonetheless reducible. That is, we may explain in terms of the microstructure of each system how it is that each aspect of the higher-level explanation is explanatory. We may, in particular, show why the RG categorisation of operators as relevant and irrelevant works. That division depends on the assumption of scale invariance, and the assumption of scale invariance is justifiable when systems are effectively scale invariant at criticality.

The anti-reductionist claim that universality is MR, and MR is essentially irreducible has been undermined by demonstrating that we may arrive at a bottom-up understanding of the common features and of what makes such features sufficient for the common behaviour.

The further argument that the use of the infinite limit imposes an irreducible divide between the higher-level and lower-level models has similarly been countered: while we move to the infinite limit in order to make the mathematics simpler, the effective scale invariance can be shown to follow from details of the particle interactions at criticality – that's what identifies the critical point and allows us to make the corresponding abstractions from scale dependent details. Provided with this bottom-up explanation, there is no further reason to claim that the infinite limit is an idealisation rather than an approximation: for we have explained from the bottom up how the system is approximately self-similar.

One upshot of this discussion is that the RG is not to be regarded as mysterious, or, somehow, as the source of physical information. It is applicable only insofar as the systems to which it is applied have the relevant properties, and their having such properties may be reductively explained.

References

Batterman, Robert W. (2000). "Multiple Realizability and Universality". In: *The British Journal for the Philosophy of Science* 51.1, pp. 115–145.

Batterman, Robert W. (2011). "Emergence, singularities, and symmetry breaking". In: *Foundations of Physics* 41, pp. 1031–1050. DOI: 10.1007/s10701-010-9493-4.

— (2016). "Philosophical Implications of Kadanoff's work on the Renormalization Group". In: *Journal of Statistical Physics (Forthcoming)*.

— (2017). "Autonomy of Theories: An Explanatory Problem". In: *Noûs*. DOI: 10.1111/nous.12191.

Binney, James J. et al. (1992). *The Theory of Critical Phenomena: an Introduction to the Renormalization Group*. Clarendon Press, Oxford.

Butterfield, Jeremy and Nazim Bouatta (2012). "Emergence and Reduction Combined in Phase Transitions". In: *AIP Conference Proceedings* 1446, pp. 383–403. DOI: 10.1063/1.4728007.

Callender, Craig and Tarun Menon (2013). "Turn and Face the Strange ... Ch-ch-changes Philosophical Questions Raised by Phase Transitions". In: *The Oxford Handbook of Philosophy of Physics*. Ed. by Robert W. Batterman. Oxford University Press, pp. 189–223.

Fisher, Michael E. (1998). "Renormalization group theory: Its basis and formulation in statistical physics". In: *Reviews of Modern Physics* 70.2, p. 653.

Franklin, Alexander (2018). "On the Renormalization Group Explanation of Universality". In: *Philosophy of Science* 85.2. DOI: 10.1086/696812.

Kathmann, Shawn M. (2006). "Understanding the chemical physics of nucleation". In: *Theoretical Chemistry Accounts* 116.1, pp. 169–182. DOI: 10.1007/s00214-005-0018-8.

Mainwood, Paul (2006). "Is More Different? Emergent Properties in Physics". PhD thesis. University of Oxford.

Morrison, Margaret (2012). "Emergent Physics and Micro-Ontology". In: *Philosophy of Science* 79.1, pp. 141–166. DOI: 10.1086/663240.

— (2014). "Complex Systems and Renormalization Group Explanations". In: *Philosophy of Science* 81.5, pp. 1144–1156. DOI: 10.1086/677904.

Norton, John D. (2012). "Approximation and Idealization: Why the Difference Matters". In: *Philosophy of Science* 79.2, pp. 207–232.

Palacios, Patricia (2017). *Phase Transitions: A Challenge for Reductionism?* URL: philsci-archive.pitt.edu/13522/.

Saatsi, Juha and Alexander Reutlinger (2018). "Taking Reductionism to the Limit: How to Rebut the Antireductionist Argument from Infinite Limits". In: *Philosophy of Science* 85.3, pp. 455–482. DOI: 10.1086/697735.

Sengers, Anneke Levelt, Robert Hocken, and Jan V. Sengers (1977). "Critical-point universality and fluids". In: *Physics Today* 30.12, pp. 42–51.

Sober, Elliott (1999). "The multiple realizability argument against reductionism". In: *Philosophy of Science*, pp. 542–564.

Wilson, Kenneth G. (1975). "The renormalization group: Critical phenomena and the Kondo problem". In: *Reviews of Modern Physics* 47.4, pp. 773–840.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. Oxford University Press.

Title: There Are No Ahistorical Theories of Function

Author: Justin Garson

Abstract: Theories of function are conventionally divided up into historical and ahistorical ones. Proponents of ahistorical theories often cite the *ahistoricity* of their accounts as a major virtue. Here, I argue that none of the mainstream "ahistorical" accounts are actually ahistorical. All of them embed, implicitly or explicitly, an appeal to history. In Boorse's goal-contribution account, history is latent in the idea of statistical-typicality. In the propensity theory, history is implicit in the idea of a species' natural habitat. In the causal role theory, history is required for making sense of dysfunction. I elaborate some consequences for the functions debate.

Keywords: Philosophy of biology; biological function; selected effects; causal role; fitness contribution

Address: Department of Philosophy, Hunter College of the City University of New York, 695 Park Ave., New York, NY 10065

Email: jgarson@hunter.cuny.edu

1. Introduction

Theories of function are conventionally divided up into two main categories, historical and ahistorical (or backwards-looking and forwards-looking). The selected effects theory (Neander 1983, 1991; Millikan 1984) is an example of a *historical* theory, but there are other historical theories, including some versions of the organizational theory (McLaughlin 2001), and the weak etiological theory (Buller 1998). *Ahistorical* theories include Boorse's goal-contribution account (1976; 1977; 2002), the propensity theory (Bigelow and Pargetter 1987), and the causal role theory (Cummins 1975; Hardcastle 2002; Craver 2001; 2013). In the 1970s and 1980s, it was common to see these two sorts of theories as competing with each other, though more recently, philosophers of biology have generally adopted a pluralistic stance, and see them as capturing different aspects of real biological usage (OMITTED). Still, the validity of the basic distinction has never been seriously challenged.

Many proponents of ahistorical theories have argued that we should accept their theories precisely *on account of* their being ahistorical. In other words, their alleged ahistoriticity is often held up as a significant virtue of their theories, and a strong reason to prefer them to historical theories (or at least a strong reason to think they capture a significant strand of ordinary biological usage). There are two arguments along these lines. The first argument appeals to bald intuition, and says that it's just obvious that functions don't always need history. One fanciful variant of this argument appeals to science fiction cases, like swamp creatures, instant lions, and randomly-generated worlds (e.g, Boorse 1976, 74; Bigelow and Pargetter 1987, 188). But one doesn't have to go as far as science fiction to find plausible cases of ahistorical functions in biology. Many philosophers have a strong intuition that, the very first time a new biological trait emerges and begins to benefit the organism, it has a *function* even if it was never selected for (e.g., Boorse 2002, 66; Bigelow and Pargetter 1987, 195; Walsh and Ariew 1996, 498). The second argument, which is closely related, appeals to ordinary biological usage, not intuition. It says that historical theories run against the way biologists ordinarily think and talk about functions. At least sometimes, when biologists attribute functions to traits, they do not *cite* or *refer to* or *think about* history or evolution (e.g., Godfrey-Smith 1993, 200; Amundson and Lauder 1994, 451; Walsh 1996, 558; Boorse 2002, 73). Hence, ahistorical theories capture important strands of real biology.

In light of the above, my thesis might come as a bit of a shock. I claim that *there are no ahistorical theories of function* – or, to put it more precisely, the mainstream versions of the allegedly ahistorical theories on the market are not actually ahistorical. If we poke and prod at those theories a bit, a historical element falls out, like contraband stashed away in a suitcase. In Boorse's version of the goal-contribution account, history is explicitly embedded in his notion of a *statistically-typical* contribution to fitness. In the propensity account, history is embedded, a little less explicitly, in the idea of a species' *natural habitat*. Finally, I claim that the only way the causal-role theorist can hope to make sense of dysfunction is to appeal to history.

If this thesis is correct – that there are no ahistorical theories of function – three consequences immediately follow. First, we need to jettison this whole way of dividing up theories of function. The distinction between etiological and non-etiological theories serves us much better, as I'll describe in the conclusion. The distinction between etiological and non-etiological theories doesn't map onto the distinction between historical and ahistorical theories; rather, *these are two ways of being historical*. Second, given that there are no ahistorical views, a good portion of the arguments that have been put forward to date for these theories (those I mentioned above) are unsound. A third consequence is that one popular way of thinking about function pluralism must fail. This sort of pluralist wishes to sort all biological usage under two main umbrella theories, the selected effects theory and the causal role theory. An argument for this sort of pluralism is that it mirrors the two main uses of "function" in biology, the historical sense and the ahistorical sense. If I'm right, this incarnation of the pluralist project can't possibly work.

Before I move on, there is one big qualification I must get out of the way. One could, just for fun, *invent* a purely ahistorical theory of function. One could assert, for example, that *all* of a trait's effects are its functions. This theory (pan-functionalism?) would be ahistorical, to be sure, since even if the world were created two seconds ago in pretty much its present form, things would still have effects, and so they'd still have functions. In fact, sometimes scientists actually *do* use the word "function" synonymously with "effect." They say things like, "climate change is a *function* of deforestation," or "poor academic performance is a *function* of malnutrition." Clearly, there are some ahistorical uses of "function." But this isn't the ordinary biological use, which the theories I cite above are trying to capture.

So, I need to amend my thesis slightly. Instead of saying that there are no ahistorical theories of function, I want to say that any theory of function that satisfies two very minimal, very traditional, and largely uncontroversial, adequacy conditions, is *also* a historical theory. First, the theory should capture some distinction between functions and accidents (the function of the nose is to help us breathe but not hold up glasses). Second, the theory should capture the possibility of malfunctioning or dysfunction. If my heart seizes up due to cardiac arrest, it's failing to perform its function or it's dysfunctional. All of the theorists I engage with in this paper purport to satisfy these two adequacy criteria, or something like them, so I'm not begging any questions by insisting on these conditions.

Here's the plan for the rest of the paper. There are five sections. After the introduction, I'll turn to Boorse's version of the goal-contribution theory, and show how it explicitly contains a historical element (Section 2). Then I'll turn to the propensity theory and show how it contains a reference to history, buried inside the idea of a trait's *natural habitat* (Section 3). I will then show how the causal-role theory, if it is to make any sense of dysfunction, must include a reference to history (Section 4). In the conclusion (Section 5), I'll reiterate the big consequences for thinking about functions and suggest a better way of dividing up theories of function.

2. Boorse's Goal-Contribution Account

Boorse's view (1976; 1977; 2002), at the most general level, is a goal-contribution account. It holds that a trait's function is just its contribution to a goal. The plausibility of this view stems from its ability to reconcile artifact and biological functions in a single theory: the function of an artifact depends on its contribution to the goal of its user; the function of a biological trait depends on its contribution to the goal of the organism or the lineage. Here, I'll focus on the subclass of functions he calls *physiological* functions.

For Boorse, the *physiological* function of a trait is its species-typical contribution to the survival and reproductive prospects of an organism (1977, 555; 2002, 72). (To be more precise, Boorse carves up species into subgroups based on age and sex; the function of a trait is its typical contribution to fitness within the members of that subgroup.) Though he doesn't define a corresponding notion of *dysfunction*, he defines a closely related notion of *disease*: a disease is simply a state that "reduces one or more functional abilities below typical efficacy."

One of Boorse's arguments for the superiority of his theory over Wright's (1973) etiological approach, and the selected effects theory of Millikan (1984) and Neander (1983), is that his approach *makes no reference to history*. He advances two arguments for the value of this ahistorical approach; one appeals to ordinary biological usage, and the other appeals to intuition. First, he says, the goal-contribution account fits ordinary biological usage: "in talking of physiological functions, they [that is, pre-Darwinian biologists] did not mean to be making historical claims at all. They were simply describing the organization of a species as they found it" (1976, 74). The same is true of current physiologists, who have "*no thought* of explaining [a trait's] history" when they assign functions to them (Boorse 2002, 73, emphasis mine). All historical theories of function simply miss how physiologists have always used the word "function." His second argument appeals to intuition. He says that intuition revolts against putting history into functions, as attested to by his instant lions case. If the lion species sprang into existence by "unparalleled saltation," one would *not* say that the parts of lions don't have functions (ibid.; also see Boorse 2002, 75). Again, functions can't be historical.

Neander (1991, 182) raised a now-famous objection against Boorse; she pointed out that Boorse's view, as it stands, can't make sense of pandemic disease: "dysfunction can become widespread within a population…A statistical definition of biological norms implies that when a trait standardly fails to perform its function, its function ceases to be its function; so that if enough of us are stricken with disease (roughly, are dysfunctional) we cease to be diseased, which is nonsense." Pandemic diseases, moreover, don't just occupy the realm of science fiction, as in P. D. James' *The Children of Men*. UV radiation poisoning in anurans is a good example of pandemic dysfunction. Sadly, climate change might create many more pandemic dysfunctions very soon. A good theory of function should at least allow for the *conceptual* possibility that all, or most, tokens of a certain trait in a certain species are dysfunctional (or as Boorse prefers, "diseased").

Intriguingly, Boorse doesn't deny the possibility of pandemic disease. Instead, he says that in order to make sense of pandemic disease, one has to appreciate function's

historical depth. Specifically, he says that when we consider what is "statistically typical" for a trait, we cannot just look at what is typical right now. Rather, we have to consider what is typical within a long slice of time that extends far back into the past: "Obviously, some of the species' history must be included in what is species-typical. If the whole earth went dark for two days and most human beings could not see anything, it would be absurd to say that vision ceased to be a normal function of the human eye (2002, 99)." He tells us that this time-slice should be longer than "a lifetime or two," and might include "millennia."

This is an extraordinary admission, given that much of Boorse's core argument *for* his view was propped up on the claim that both biology and intuition need purely ahistorical functions, uncluttered by history. His admission implies that two of his key arguments for the view (cited above), are unsound. First, by his own admission, it's not the case that biologists don't refer to history; implicitly, when they talk about what's statistically-typical, they *are* talking about history. Second, regardless of whether or not intuition supports ahistorical functions, Boorse's theory doesn't. It's just not true, on Boorse's account, that if lions popped into being from an unparalleled saltation, their parts and processes would have functions. They wouldn't, since they don't have the right history (or to be more precise, they have no history at all). True, Boorse's history isn't the same *kind* of history that features in the selected effects theory, since it doesn't refer specifically to etiology, but it's still history, and so his arguments that appeal to the ahistoricity of his theory don't work.

3. The Propensity Theory

Bigelow and Pargetter (1987) also developed an influential "ahistorical" theory of function, the propensity theory. They reject the selected effects theory (and etiological accounts more generally) because the selected effects theory gets the *modality* of functions wrong. In other words, the statement, "functions are selected effects," if true, is contingently true; it might be true on the actual world, but there are possible worlds at which it's false. To illustrate the point, they ask us to consider a world that is pretty much the same as ours except that it randomly popped into being five minutes ago. On that world, they claim, there would still be functions, just no selected effects (188): "we have the intuition that the concept of biological function…[is] not thus contingent upon the acceptance of the theory of evolution by natural selection." This consideration prompts the need for an ahistorical theory.

For Bigelow and Pargetter, functions are propensities, or probabilistic dispositions. We might quibble over what exactly dispositions are, but any good definition will cite three parts: structure, environment, and behavior. Consider the solubility of salt. There is a *structure*, namely, the polar molecular structure composed of sodium and chloride; there is an *environment*, namely, water; there is a *behavior*, namely, dissolving. When we say that salt is disposed to dissolve in water, we're saying that, if you were to take this structure, and put it in this environment, it would perform this behavior.

Functions, too, are dispositions. Consider "the function of the heart is to circulate blood." For this statement to be true, there must be a structure (the heart, embedded the right way in the circulatory system), an environment (which they call the creature's *natural habitat*), and a behavior (conferring a fitness boost on the organism). If one were to put the structure in its natural habitat, it would increase the fitness of the organism (relative, I suppose, to creatures without hearts). The crucial distinction between their view and Boorse's is that in their view, a trait's function doesn't depend on actual frequencies of performance. A trait needn't have an actual track record of boosting fitness to have a function; a mere propensity will do.

This raises the thorny question of what a creature's *natural habitat* is. For they're clear that a creature's natural habitat isn't just any environment the creature happens to find itself in. Unfortunately, they refuse to define this crucial notion; instead, they brush it off as vague, but unproblematically so: "there may be room for disagreement about what counts as a creature's 'natural habitat;' but this sort of variable parameter is a common feature of many useful scientific concepts" (192). But one could at least form the suspicion that if one analyzed this unproblematically vague notion, one would find some reference to history tucked away inside of it.

This suspicion is confirmed in the very next paragraph. There, they tell us that, if a creature's environment were to change very suddenly, then "natural habitat" will still refer to the *old* environment, and not the *new* one (ibid). There's a time lag built into the very idea of a natural habitat. So, for example, if climate change melts enough Arctic ice, then, at least for a time, the polar bear's natural habitat (and by extension, the natural habitat of the trait itself, namely, their thick, water-repellant fur) is the icy habitat of yore and not the contemporary, denuded one. They take that as given, and I agree.

But why would this be? What *makes it the case* that this is true, namely, that in cases of rapid habitat change, "natural habitat," at least for a time, refers to the old environment and not the new one? What makes it true, I suspect, is that the idea of a natural habitat is an intrinsically historical notion. It's something like the environment within which the organism recently survived and thrived. And if that's not what a natural habitat is, I would like to know what it is *such that*, if a creature's actual habitat shifts suddenly, the natural habitat is still the old one. Just because a concept is vague around the edges, that doesn't exempt one from the obligation to give some sort of analysis.

Hence, I conclude that, contrary to rumor, the propensity theory is not an ahistorical theory, or not demonstrably so. But if that's right, they lose one of the main virtues of the view, which is to get the modality of functions right. To be fair, there's still a sense in which their view *is* ahistorical. What they can do, that the selected effects theorist can't, is to attribute functions to novel traits – so long as that novel trait belongs to the members of a species that has been around long enough to have a natural habitat. Suppose a gene mutation confers a benefit on an organism, say, pesticide resistance on a flour beetle. I suppose they can say that, at the very moment at which it first confers that benefit, the gene mutation has a function, namely, to make the beetle withstand a certain pesticide. This result, they claim, is "intuitively comfortable" (195). But they can say that only

because flour beetles themselves have a history, and so we can talk meaningfully about their natural habitats. Moreover, I think they'll still have a very hard time dealing with dysfunction (Neander 1991, 183), as I hope to show in the next section. Finally, I think there are good theory-neutral reasons for saying that beneficial traits, on their very first appearance, don't have functions, but rather, whatever benefit they bring is an accident. But I won't argue for that here (see OMITTED).

4. The Causal Role Theory

What about the causal role theory of function? This appears to be a purely ahistorical view. The causal role theory says, roughly, that the function of a *component* of a system consists in its contribution, in tandem with the other components, to a system-level capacity of interest (Cummins 1975; Craver 2001; Hardcastle 2002). Craver (2001; 2013) helpfully elaborates this view by specifying that the part in question must be a component of a *mechanism*. All of the basic ingredients of this theory are ahistorical: capacities, components, organization, hierarchy, interests. Even if the world were created five minutes ago, in pretty much its present form, things would still have causal role functions.

The problem enters when we think about dysfunction. Cummins (1975, 758) insisted that functions are dispositions, or capacities: "…to attribute a function to something is, in part, to attribute a disposition to it." The function of a trait *token*, then, consists in its capacity to contribute to a system-level effect. But what if the token in question, through defect or disease, loses the capacity, and so can't contribute to the system-level effect? Then, by Cummins' analysis, it doesn't have the relevant function – so it can't dysfunction either.

Causal role theorists have, by and large, been silent about how to make sense of dysfunctions from this perspective. Almost everything they've had to say on that score, however, is consistent with the following theme: a trait *token* dysfunctions when it can't do what other trait tokens generally, or typically, do to contribute to the system-level effect of interest. Consider Godfrey-Smith (1993, 200): "Although it is not always appreciated, the distinction between function and *mal*function can be made within Cummins' framework…If a token of a component of a system is not able to do whatever it is that other tokens do, that plays a distinguished role in the explanation of the capacities of the broader system, then that token component is malfunctional." Craver (2001, 72), offers the same general line: "…the ascription of a function to a malformed or broken part is derivative upon a description of how that *type* of part (X) fits into a *type* of higher-level mechanism (S). The malformed and broken part can be identified as an X by the typical properties and activities of Xs...." This is, at root, to rely on a statistical norm for making sense of dysfunction.

This account of dysfunction, like Boorse's, stumbles when it encounters the problem of pandemic dysfunction (Neander 1991). For the modification suggested above implies that, if everyone's heart seized up at once, nobody's heart would have a function anymore, so nobody's heart would be dysfunctional. The best way to solve this problem,

and perhaps the only way, is the way Boorse took, namely, to say that the function of a trait is its typical contribution to some system effect, when what's typical is assessed over a chunk of time that stretches back into the past, for at least "a lifetime or two," and perhaps "millennia." But if causal role theorists take that line, they'd have a historical theory.

Craver (2001) and Hardcastle (2002) suggest, all too fleetingly, a different way of thinking about dysfunction, one that depends not on statistics, but on our values and goals, that is, the values and goals of people who make function attributions. Craver (2001, 72) suggests that traits dysfunction when they cannot do what people *want* them to do: "the mechanistic role of the broken part only appears against the fixed backdrop of shared assumptions about a type of mechanism within which parts of this type generally (or preferably) make important contributions." The parenthetical remark alludes to a substantially new doctrine, one that demands our full concentration. It suggests that dysfunction is a mirror of human preferences and goals, of our wishing and wanting. If my heart seizes up, it's dysfunctional, since it's not doing *what I want it to do.*

Hardcastle (2002) makes remarks along similar lines. She first says that the function of a trait - what it's "supposed to do," as she puts it - depends on the goals of the scientific discipline that makes the investigation: "The teleological goal for some trait…depends upon the discipline generating the inquiry" (153). The palmomental reflex causes a chin twitch when you stroke an infant's palm; it's just an accident of cortical wiring with no deep evolutionary rationale. Still, she says, it has the *function* of indicating the state of brain development in infants, because that's how biomedical researches use it. She then says that something malfunctions just when it cannot do what it's supposed to do (152). The palmomental reflex malfunctions when it can't indicate the state of brain development. Simply put, dysfunction happens when a trait can't do what we want.

But dysfunctions cannot be reduced to preferences in any straightforward way; this is a point that's been taken for decades (e.g., Boorse 1977, 544; Wakefield 1992, 372), for reasons that scarcely need to be rehearsed. I'd prefer not to need sleep and water; I'd prefer if nobody had to go through the pain of childbirth or teething, either. But none of those things are diseases or dysfunctions. For that matter, I'd prefer if my hands were equipped with retractable adamantium claws. The fact that my hands can't do what I want them to do doesn't make them dysfunctional. If one really wanted to run with this value-centered line about dysfunction, one would *at least* have to add that, in order for a trait to dysfunction, it's not enough that it doesn't do what I prefer, but I must also have a *reasonable expectation* that it *should* act in the way that I prefer. But what could possibly ground a *reasonable expectation* that my hand (say) work in a certain way? Only this: that hands usually *do* work in the preferred way. But then we're back to statistical norms, and long historical slices of time. This value analysis of dysfunction isn't a contender to a statistical analysis; instead, the former presupposes the latter.

I've walked through three allegedly ahistorical theories of function, and shown that none of them are purely ahistorical; they're tainted with history. The conclusion will say what we should do next.

5. Conclusion

There are no ahistorical theories of function, at least among those that are usually put forward as ahistorical. The first, Boorse's goal-contribution theory, explicitly refers to what is statistically typical for a trait, where what's typical is assessed over a long historical period of time. The second, the propensity theory, refers to the creature's natural habitat, which is implicitly historical. And the third, the causal role theory, can't hope to make sense of dysfunction (or so I argue) without appealing to a statistical norm, and thereby (following Boorse) to history. *No* theory of function will give functions to the parts of swamp creatures, instant lions, or anything on worlds that are similar to ours except for being randomly generated five minutes ago. The propensity theory, at least, can give functions to novel traits as soon as those traits begin benefiting their bearers, as long as the population in which the traits emerge has been around for long enough to have something like a natural habitat. But even that theory will probably encounter problems when it comes to making sense of dysfunction, though I haven't pushed that line in any detail here.

Three immediate consequences follow from this fact. The first is that we should stop dividing up theories of function in terms of historical and ahistorical. The second is that many of the main arguments for the allegedly ahistorical theories are unsound. Third, one popular form of pluralism, which says that there are two main theories of function, corresponding to historical and ahistorical uses of "function" in biology, is untenable.

But if we can't rely on the historical/ahistorical distinction as a way of dividing up functions, how should we talk about them? I think it's best to divide them up into etiological and non-etiological (as theorists are sometimes wont to do anyway). But there's a crucial clarification in order: to say a theory is etiological isn't *just* to say it's historical. It's to say that the theory deals specifically with causal history. The theory purports to capture the sense in which, when we attribute a function to a trait, we're trying to give a causal explanation for why the trait exists. Most other theories of function are non-etiological, in that they do not purport to explain, in a causal sense of "explain," why the trait exists. But they're still historical.

There's a twist to this story. I think there *are* ahistorical theories of function. Consider that climate change is a function of deforestation, poor academic performance is a function of malnutrition, and wildlife habitat is a function of soil. These notions are *ahistorical* through and through. "Function," in this context, means little more than "effect," and perhaps (as in the last of the three examples) "helpful effect." But this tepid sense of function isn't going to sustain a distinction between function and accident, nor will it give us any sense of dysfunction. This is the sort of "function" that Bock and von Wahlert (1965, 274) were getting at when they equated functions with "all physical and chemical properties arising from [the trait's] form." It's also the sort of "function" that Neander (2017) describes in her recent discussion of "minimal functions." But the proponents of the allegedly ahistorical theories want functions to do much more than that. They are trying to capture the ordinary biological sense (or *an* ordinary biological sense)

of "function," where functions differ from accidents and sometimes things dysfunction. Unfortunately, they can't have what they want.

References

Amundson, R., and G. V. Lauder. 1994. Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy* 9: 443-469.

Bigelow, J., and Pargetter, R. 1987. Functions. *Journal of Philosophy* 84: 181-196.

Bock, W. J., and von Wahlert, G. 1965. Adaptation and the form-function complex. *Evolution* 19: 269-299.

Boorse, C. 1976. Wright on functions. *Philosophical Review* 85: 70-86.

Boorse, C. 1977. Health as a theoretical concept. *Philosophy of Science* 44: 542- 573.

Boorse, C. 2002. A rebuttal on functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M. Perlman, 63-112. Oxford: Oxford University Press.

Buller, D. J. 1998. Etiological theories of function: A geographical survey. *Biology and Philosophy* 13: 505-527.

Craver, C. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Craver, C. 2013. Functions and mechanisms: A perspectivalist view. In *Function: Selection and Mechanisms*, ed. P. Huneman, 133-158. Dordrecht: Springer.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72: 741–765.

Godfrey-Smith, P. 1993. Functions: Consensus without unity. *Pacific Philosophical Quarterly* 74: 196-208.

Hardcastle, V.G. 2002. On the normativity of functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M Perlman, 144-156. Oxford: Oxford University Press.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Neander, K. 1983. *Abnormal Psychobiology*. Dissertation, La Trobe.

Neander, K. 1991. Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.

Neander, K. 2017. Functional analysis and the species design. *Synthese* 194: 1147-1168.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47: 373–388.

Walsh, D.M. 1996. Fitness and function. *British Journal for the Philosophy of Science* 47: 553-574.

Walsh, D. M., and A. Ariew. 1996. A taxonomy of functions. *Canadian Journal of Philosophy* 26: 493-514.

Wright, L. 1973. Functions. *Philosophical Review* 82: 139-168.

What do molecular biologists mean when they say 'structure determines function'?

Gregor P. Greslehner*

University of Salzburg & ERC IDEM, ImmunoConcept, CNRS/University of Bordeaux

October 2018

Abstract

'Structure' and 'function' are both ambiguous terms. Discriminating different meanings of these terms sheds light on research and explanatory practice in molecular biology, as well as clarifying central theoretical concepts in the life sciences like the sequence–structure–function relationship and its corresponding scientific "dogmas".

The overall project is to answer three questions, primarily with respect to proteins: (1) What is structure? (2) What is function? (3) What is the relation between structure and function?

The results of addressing these questions lead to an answer to the title question, what the statement 'structure determines function' means.

*Email: gregor.greslehner@gmail.com

Keywords: philosophy of biology, molecular biology, protein structure, biological function, scientific practice

1 Introduction

'Structure' and 'function' are abundantly used terms in biological findings. Frequently, the conjunct phrase 'structure *and* function' or the directional phrase '*from* structure *to* function' is to be found, indicating that there is a special relation connecting these two concepts. The strongest form of this relation is found in the frequent statement that 'structure *determines* function'. One could easily list several hundreds of references containing such phrases. However, in order not to blow up the references section, I will refrain from doing so. Suffice it to say that biologists make highly prominent use of these concepts in describing their research—molecular biologists, in particular. In this paper, I attempt to clarify these concepts, address their relation, and discuss the role they play in molecular biology's explanatory practice. While these issues can be addressed for many different biological entities on different levels of organization, I restrict the discussion primarily to proteins.

What do biologists refer to when they use this phrase? Is there a particular scientific program or strategy behind the slogan 'structure determines function'? Despite the frequent use of this phrase and the concepts to which it refers, a rigorous analysis is missing. Thus, a philosophical clarification would be a valuable contribution to the conceptual foundations of biology. One such fundamental concept is the sequence–structure–function relationship. "The relationships between sequence, structure, biochemical function and biological role are extremely ill-defined and scant

high quality data are available to allow us to analyse them." (Sadowski and Jones, 2009, 360)

In this paper, I attempt to close this gap by developing an explication of both concepts of *structure* and *function* as they are used in biological practice and discussing which relation holds between them. The third component in this "trinity of molecular biology"—sequence—is the least in need of explication. The standard textbook view holds that sequence determines structure, and structure determines function. I will focus on the second relation.

Without reviewing the rich history of these concepts throughout biology at this point, it is worth noting that functionality and form or structure were thought to be intimately linked from early on. In the early days of biology at the macroscale, the structures had to be observed with the naked eye. Thus, the first examples about the form of bodies or their parts and their functions can be found in physiology and anatomy, for example Harvey's notion of the heart's function to pump blood. From the scale of physiology to the molecular scale, structure and function are closely related. What exactly links these two concepts? Is it a determination relation? And if so, which one is determining the other?

With the invention of microscopes and later the emergence of molecular biology, the structures and functions under consideration shifted from macroscopic entities to individual molecules. In fact, molecular biology put the three-dimensional shape of molecules center stage for explaining biological phenomena. This is the focus of this paper. In particular, the discussion will be confined to the structure and function of *proteins*—with special emphasis on the question whether the former determines the latter.

2 The ambiguity of 'structure'

In a first approximation, 'structure' and 'function' could be interpreted as the most general or neutral way of describing what molecular biologists are doing in their research and what their findings are about. These include mainly the three-dimensional shapes of molecules (or larger cellular structures) and the activities (functions) theses molecules perform in living cells, biochemical pathways, chemical reactions, or just individual steps in such mechanisms. The ultimate aim is to explain biological phenomena with molecular mechanisms, whose entities can be described in physical and chemical terms. The structure of molecules can be described in terms of physics and chemistry—function, however, is a concept that does not appear in physics or chemistry. Let's start by taking a closer look at the notion of structure.

'Structure' is an ambiguous term. Applied to proteins, there is the usual nomenclature of *primary structure* (i.e., a protein's amino acid sequence), *secondary structure* (i.e., common structural motifs like α-helices and β-sheets), *tertiary structure* (i.e., the three-dimensional shape of a single folded amino acid chain), and *quaternary structure* (i.e., the final assembly of a protein if it consists of more than one amino acid chain). Other structurally important components are post-translational modifications and prosthetic groups which are not part of its amino acid composition. All these notions of structure have in common that they are about the molecular composition and shape of a molecule. One meaning of 'structure' denotes the sequence of a polymer, the other meaning is about the three-dimensional shape of a molecule. As will be discussed below, another important ambiguity of 'structure' allows to denote the organization of an interaction network. That leaves us with three different meanings of 'structure':

(1) the sequence of a polymer, (2) the three-dimensional shape of a molecule, and (3) the network organization of several biological entities.

While meanings (2) and (3) are candidates for being functional entities, structure as sequence (1) rather relates the sequences of different polymers (DNA, RNA, and proteins) and also plays a central role in determining the three-dimensional shape of a molecule, structure (2). The primary structure of a protein is just the sequence of amino acids that are put together to form a polypeptide. This amino acid sequence is determined by the corresponding protein-coding gene, which is first transcribed into mRNA and then translated into protein by the ribosome. This scheme is known as the "central dogma of molecular biology":

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$$

The arrows might be interpreted as determination relations. The textbook view of protein structure and function proceeds as follows:

$$\text{nucleotide sequence} \rightarrow \text{amino acid sequence} \rightarrow \text{protein structure} \rightarrow \text{protein function}$$

Strong evidence supporting the claim that the three-dimensional shape of a protein is determined by the sequence of amino acids alone was provided by the experiments of Christian Anfinsen, showing that ribonuclease could, after treatment with denaturing conditions, regain its form and function (Anfinsen et al., 1961). Later, Merrifield showed that an *in vitro* synthesized sequence of amino acids can carry out the enzymatic activity of ribonuclease, thus gaining its functional form without the aid of any other cellular component (Gutte and Merrifield, 1971). From this and similar experiments, Anfinsen

built general rules of protein folding as a global energy minimum which depends solely on the sequence of amino acids (Anfinsen, 1973). This view is known as "Anfinsen's dogma".

In 1958, John Kendrew's lab determined the first actual three-dimensional form of a protein, myoglobin (Kendrew et al., 1958). The predominant technique to determine protein structures is still X-ray crystallography (Mitchell and Gronenborn, 2017). Other techniques include nuclear magnetic resonance, cryogenic electron microscopy, and atomic force microscopy. X-ray structures in particular have been supporting the view that there is a unique rigid shape—the protein's native, functional state—which would be necessary and sufficient for a protein to carry out its biological function.

To make a long story short, the relation between nucleotide sequence and amino acid sequence has been generally confirmed (although there are much more complicated mechanisms to it, e.g., splicing). However, the part concerning the protein shape and function proves to be much more problematic. That poses a challenge to what Michel Morange calls "the protein side of the central dogma" (Morange, 2006).

To get from amino acid sequence to three-dimensional structure is known as the *protein folding problem*. As the term 'problem' suggests, it poses a serious challenge and remains unsolved to this day. Even though knowledge-based techniques to predict protein structures from their sequence have become impressively sophisticated, successful, and reliable, there are good reasons to suspect that the protein problem might remain unsolved in principle—if the aim is to predict protein folding based on chemical and physical principles only.

Every two years the best prediction tools are tested in a contest, the Critical Assessment of protein Structure Prediction (CASP). Based on experimentally determined structures which are only published after the participants of the contest have

submitted their predictions, the predictions are then compared to the experimental structure. A similar contest for predicting the functions of proteins exists (Critical Assessment of Functional Annotation, CAFA), although it is much less developed. But what is function in the first place?

3 The ambiguity of 'function'

'Function' is also an ambiguous term (Millikan, 1989)—even more so than 'structure'. There is a rich history of debates surrounding different notions of function. The term 'function' has a long tradition in biology and its philosophy (Allen, 2009). Starting with Aristotle, activities in biology were interpreted to *have a purpose*, to be goal-directed (teleological). The standard example is that the heart's function is to pump blood. That the heart also produces noise is not considered to be functional. Classic accounts of function have been predominantly trying to capture the teleological aspect, for example (Wright, 1973). However, intentionality is a problematic notion in biology. In another important account, Robert Cummins (1975) stressed the importance of a component's contribution to the system in which it is contained, rather than why natural selection has favored a certain trait. Although it makes sense in evolutionary biology to have an account of function that captures the evolutionary developments, molecular biology and protein science operate with a different notion of function, i.e., mainly biochemical activity. There seem to be two entirely different questions: What is a structure doing? And how did this structure evolve to do what it does?

Arno Wouters distinguishes four notions of biological function (Wouters, 2003): (1) (mere) activity, (2) biological role, (3) biological advantage, and (4) selected effect.

The last two are issues of evolutionary biology, whereas the former two fall within the molecular biologist's domain. If function is to be determined by a molecule's three-dimensional shape or organization network, only (1) and (2) seem to be the proper reading of 'function' in this context.

Which entities have functions within living organisms? Depending on the level of organization at which one is operating, one could give a different answer: molecules, organelles, cells, tissues, organisms, individuals, populations, ecosystems. The most prevalent candidates in molecular biology are certainly DNA and proteins, although lipids and other biomolecules play important roles in life processes, too.

Traditionally, functions have been attributed to entire genes ("one gene—one enzyme hypothesis"). These views are related to the genetic determinsim view of having a gene for every trait, in which every gene has a function. However, the primary functional units inside a cell are arguably its proteins. Their biochemical activities and biological roles depend crucially on their three-dimensional shapes and network organization, respectively.

One has also to take into account more abstract functional entities, i.e., network modules. These are also called 'structures' but do not refer to the shape of molecules. Its functions ought to be considered as Wouter's second notion (biological role), rather than biochemical activity. "Current 'systems' thinking attributes primary functional significance to the collective properties of molecular networks rather than to the individual properties of component molecules" (Shapiro, 2011, 129). "[A] discrete biological function can only rarely be attributed to an individual molecule [...]. In contrast, most biological functions arise from interactions among many components." (Hartwell et al., 1999, C47). Thus, we can attribute functions as biochemical activities to

individual molecules, wheres systems functions (biological roles) are attributed to organizational structures:

> "Finding a sequence motif (e.g., a kinase domain) in a new protein sheds light on its biochemical function; similarly, finding a network motif in a new network may help explain what systems-level function the network performs, and how it performs it." (Alon, 2003, 1867)

4 Does structure determine function?

Having distinguished between three notions of 'structure' and two notions of 'function', what about the statement 'structure determines function'? Is—in any of its different readings—a certain structure necessary or sufficient for a certain function?

The common textbook view according to Anfinsen has a clear answer: "the central dogma of structural biology is that a folded protein structure is necessary for biological function" (Wright and Dyson, 1999, 322). On first glance, it might appear plausible to assume that a particular structure (understood as molecular shape) is a necessary condition for the proper function of a biological structure (i.e., its biochemical activity). Loss of function is often associated with a loss of the three-dimensional shape of individual proteins. On the other hand, to go for the "sufficient" direction, changes in structures often lead to a decrease in functionality, up to a complete loss. Many diseases for which there are known molecular causes give support to this view. Often it is alterations in the sequence of DNA that result in changed protein shapes that lead to a functionality defect of the organism, which is the definition of a "molecular disease". Alterations of a protein's three-dimensional shape, however, do not necessarily lead to

loss of function. In many cases, changes are "silent", i.e., they don't cause any alteration in phenotype. In rare events, changes might even turn out to be "improvements", which is the driving force of evolutionary development.

However, evidence has been found in the recent years that a significant portion of proteins are intrinsically unstructured in order to be functional, see for example (Forman-Kay and Mittag, 2013). Does the discovery of intrinsically unstructured proteins challenge the relation between structure and function? "[D]isorder aficionados are calling for a complete reassessment of the structure-function paradigm" (Chouard, 2011, 151). Some protein domains fold only upon binding to a suitable target. Others, however, seem to never have an ordered state at all—they remain unstructured even in their functional state.

That a high similarity in sequence does not guarantee a similarity in structure or function has been shown by the Paracelsus Challenge: "a one-time prize of $1000, to be awarded to the first individual or group that successfully transforms one globular protein's conformation into another by changing no more than half the sequence" (Rose and Creamer, 1994, 3). One recent answer to this challenge resulted in the synthesis of two proteins which have 88% sequence identity but a different structure and a different function (Alexander et al., 2007).

Contrary to the view described above, the generalization that a stable three-dimensional structure is necessary or sufficient for a particular function does not hold. It remains true, however, that there is an intimate correlation between structure and function. Prediction tools based on this view are a powerful tool. An attempt to systematically predict the structure and function of proteins based on their amino acid sequence can be found, for example, in (Roy et al., 2010).

To complicate the picture, codon usage is also important: Zhou et al. (2013) have shown that the FRQ protein, which is involved in the circadian clock, is using non-optimal codons, thus translation speed is not optimal. After experimentally optimizing codon usage, the resulting protein—which has the exact same amino acid sequence—folds differently and is no longer functional. This shows that amino acid sequence by itself is not sufficient to determine the three-dimensional structure, let alone its function. In addition to the correct sequence, the folding process has to take place in a certain way which is influenced by the usage of codons and thus the availability of tRNAs, which influences the speed at which the ribosome can proceed translation. Usage of non-optimal codons gives the nascent polypeptide chain some time for the segments that have already been translated to fold in a certain conformation. If translation is too fast, certain intermediate folds which are necessary to reach the final functional conformation can be lost.

Another idea to keep in mind is that evolution operates pragmatically: structures are not the target of selection, functions are. Structures are being re-used for novel functions—there are many biological examples.

If structure does not *determine* function, if a particular structure (in any of its three meanings) is neither necessary nor sufficient for a particular function (in any of its two meanings), may there be another way in which structure and function are related? Perhaps there is a less stringent relationship? I will argue for a supervenience relation (McLaughlin and Bennett, 2018). But before developing this account, we need to clarify which notions of 'structure' and 'function' to use to capture actual scientific practice in molecular biology.

In order to speak about biological functions, a reglemented vocabulary is needed.

The most successful of these is gene ontology (GO) (Ashburner et al., 2000). Fascinating correlation analysis between three-dimensional protein structures from the Protein Data Bank (PDB) and GO terms can be found, for example, in (Hvidsten et al., 2009) and (Pal and Eisenberg, 2005).

According to the textbook picture, there is a linear chain of determination, leading from nucleotide sequences in the DNA via transcription to the nucleotide sequence of RNA, which leads via translation to the amino acid sequence of proteins. The sequence of amino acids, in turn, determines the three-dimensional structure of the protein, whose function, again, is determined by its structure. Given transitivity of this determination relation, one would only need to know the genomic sequence in order to have a complete picture ("blue print") of the functional organism. That is the "holy grail of molecular biology". And like the quest for the holy grail, it is doomed to fail. A strict determination relation does not even hold between the individual pairs.

The reason why the simplified scheme above is still part of the current research "paradigm" lies, on the one hand, in its scientific success: genomics and proteomics have provided unimaginable insights. On the other hand, it fits the mechanistic, reductionistic narrative that has been fashionable in molecular biology. Today, systems biology claims to provide a "holistic" alternative (Green, 2017).

But even without such a strict determination relation between structure and function, both concepts are central to explaining molecular mechanisms in research practice.

In order to understand why molecular biologists explain mechanisms with reference to structure and function, we need to understand what these concepts denote. In a first approximation, molecular biologists analyze a phenomenon by identifying its components that are responsible for the phenomenon in question. These components are the

structures that perform certain biochemical activities, which collectively bring about the phenomenon (biological role). The way in which these entities and their activities are organized is a different meaning of 'structure' which is as important in a mechanistic explanation as individual molecular structures are.

> "Despite the lack of an overarching theory, a Newtonian or quantum mechanics of its very own, molecular biology has become a unifying discipline in virtue of the powers of its techniques, its ability to extrapolate from the molecular to higher levels, and its synthesis of problems of form and function at the molecular level. This synthesis of form and function is a central, ill-understood, and historically important feature of molecular biology." (Burian, 1996, 68)

The ambiguity of the terms 'structure' and 'function' might be useful, for it can be applied to a broad variety of biological research strategies and activities. But, on the other hand, using the term same for different things causes confusion, and the use of metaphorical language might be obscuring certain features and difficulties with this approach.

More recent and thriving approaches in the life sciences have moved beyond the idea that there is a determination relation between structure and function and that by knowing the structure of a protein one could predict its biological function. Today's research in molecular biology is more centered around the *organizational structure* of biological mechanisms. In this way, the ambiguity of the term 'structure' suits to uphold the research slogan, since it can also be applied in a broader sense here than just molecular shapes. The organization of biological systems is the domain of the relatively

new discipline systems biology.

The three-dimensional shape is often a detail that does not contribute to the understanding of a mechanism, but to the contrary would only confuse the mechanistic picture which requires a certain level of abstraction in order to be comprehensive.

But still, how exactly do we get from molecular structures and their (structured) activities to biochemical activities and biological functions? That there might not exist a straightforward mapping from molecular shapes to their biochemical and biological function had been anticipated in the early days of molecular biology:

> "It [molecular biology] is concerned particularly with the *forms* of biological molecules, and with the evolution, exploitation and ramification of these forms in the ascent to higher and higher levels of organization. Molecular biology is predominantly three-dimensional and structural—which does not mean, however, that it is merely a refinement of morphology. It must of necessity enquire at the same time into genesis and function." (Astbury, 1952, 3, original emphasis)

Taking up Francis Crick's remark that "folding is simply a function of the order of the amino acids" (Crick 1958, 144), Morange comments that it is "obviously not a *simple* function" (Morange, 2006, 522). And he observes a semantic change in the meaning of 'function':

> "For Francis Crick, function meant the application of simple rules and principles. For specialists today, function is the result of a complex evolution [...] This shift in the meaning of a word is more than anecdotal. It reflects an active ongoing transformation of biology [...] The mechanistic models of

> molecular biology are no longer considered sufficient to explain the structures and functions of organisms. They have to be complemented and allied with evolutionary explanations" (Morange, 2006, 522).

In order to explain biological phenomena, there is no determination relation that would allow us to track everything down to the chemical and physical properties of proteins, let alone the nucleotide sequences of DNA. Of course, all these issues are relevant to the topic of reduction:

> "if [...] regulatory networks turn out to be crucial to explaining development (and evolution [...]), the reductionist interpretation *may* be in trouble. If network-based explanations are ubiquitous, it is quite likely that what will often bear the explanatory weight in such explanations is the topology of the network rather than the specific entities of which it is composed. [...] How topological an explanation is becomes a matter of degree: the more an explanation depends on individual properties of a vertex, the closer an explanation comes to traditional reduction. The components matter more than the structure. Conversely, the more an explanation is independent of individual properties of a vertex, the less reductionist it becomes." (Sarkar, 2008, 68, original emphasis)

5 Conclusion

Both terms, 'structure' and 'function', are highly ambiguous. So is the widely used conjunct phrase of 'structure and function' that is ubiquitous in biology, as well as the

even stonger claim 'structure determines function'. Perhaps this is why it can be used in many different contexts and for many different explanatory aims in biology. Although providing a certain framework of generality, I argue that a clarification of these concepts is beneficial—for conceptual and philosophical considerations, as well as for the way biologists think about the grand schemes like the "central dogma". Ideally, such an account would also have practical implications and benefit current biological research.

To sum up the results of my analysis, in molecular biology's explanatory practice, 'structure' may refer to:

1. the sequence of polymers,

2. the three-dimensional shape of molecules (or their parts), and

3. the way biological entities are organized.

Of course, different aspects of this distinction play different roles in the explanatory practice with respect to molecular mechanisms. The detailed shape of the interacting molecules is neither necessary nor sufficient for understanding its activities (although correlations are valuable prediction tools before doing experiments in the lab).

The ambiguity of the term 'function' depends on whether the explanation aims at answering the question how a mechanism works or how it came to work that way. Even in the first case one has to distinguish between:

1. the biochemical activity of individual components, and

2. the biological role of network structures.

Whereas biochemical activities of proteins can often be successfully predicted by homology modeling from known molecular shapes, the biological role is rarely an

intrinsic property of an isolated molecule. Rather, the biological role is the mechanistic result of an interaction network of several dynamically interacting molecules.

By comparing the combinatorial possibilities of the different meanings of 'structure' and 'function', a determination relation does not hold between any of them. Instead, I propose a supervenience relation: between the three-dimensional shapes of protein domains and their biochemical activities, and between interaction networks and their biological role. According to my analysis, this is what molecular biologist mean when they say 'structure determines function'.

References

Alexander, P. A., Y. He, Y. Chen, J. Orban, and P. N. Bryan (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences 104*(29), 11963–11968. doi:10.1073/pnas.0700922104.

Allen, C. (2009). Teleological notions in biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009 ed.). `http://plato.stanford.edu/archives/win2009/entries/teleology-biology/`.

Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science 301*(5641), 1866–1867. doi:10.1126/science.1089072.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science 181*(4096), 223–230. doi:10.1126/science.181.4096.223.

Anfinsen, C. B., E. Haber, M. Sela, and F. H. White, Jr (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences 47*(9), 1309–1314.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. *Nature Genetics 25*, 25–29. doi:10.1038/75556.

Astbury, W. T. (1952). Adventures in molecular biology. In *The Harvey Lectures. Delivered under the auspices of the Harvey Society of New York. 1950–51*, pp. 3–44. Charles C Thomas.

Burian, R. M. (1996). Underappreciated pathways toward molecular genetics as illustrated by Jean Brachet's cytochemical embryology. In S. Sarkar (Ed.), *The Philosophy and History of Molecular Biology: New Perspectives*, pp. 67–85. Kluwer Academic Publishers.

Chouard, T. (2011). Breaking the protein rules. *Nature 471*, 151–153. doi:10.1038/471151a.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy 72*(20), 741–765. doi:10.2307/2024640.

Forman-Kay, J. D. and T. Mittag (2013). From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure 21*(9), 1492–1499. doi:10.1016/j.str.2013.08.001.

Green, S. (2017). Philosophy of systems and synthetic biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 ed.). `https://plato.stanford.edu/archives/sum2017/entries/systems-synthetic-biology/`.

Gutte, B. and R. B. Merrifield (1971). The synthesis of ribonuclease A. *Journal of Biological Chemistry 246*, 1922–1941.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray (1999). From molecular to modular cell biology. *Nature 402*(6761 Suppl.), C47–C52. doi:10.1038/35011540.

Hvidsten, T. R., A. Lægreid, A. Kryshtafovych, G. Andersson, K. Fidelis, and J. Komorowski (2009). A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLoS ONE 4*(7), e6266. doi:10.1371/journal.pone.0006266.

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, and H. Wyckoff (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature 181*(4610), 662–666. doi:10.1038/181662a0.

McLaughlin, B. and K. Bennett (2018). Supervenience. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). `https://plato.stanford.edu/archives/spr2018/entries/supervenience/`.

Millikan, R. G. (1989). An ambiguity in the notion "function". *Biology and Philosophy 4*, 172–176. doi:10.1007/BF00127747.

Mitchell, S. D. and A. M. Gronenborn (2017). After fifty years, why are protein X-ray

crystallographers still in business? *The British Journal for the Philosophy of Science 68*(31), 703–723. doi:10.1093/bjps/axv051.

Morange, M. (2006). The protein side of the central dogma: Permanence and change. *History and Philosophy of the Life Sciences 28*(4), 513–524.

Pal, D. and D. Eisenberg (2005). Inference of protein function from protein structure. *Structure 13*, 121–130. doi:10.1016/j.str.2004.10.015.

Rose, G. D. and T. P. Creamer (1994). Protein folding: Predicting predicting. *PROTEINS: Structure, Function, and Genetics 19*, 1–3.

Roy, A., A. Kucukural, and Y. Zhang (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols 5*(4), 725–738. doi:10.1038/nprot.2010.5.

Sadowski, M. and D. T. Jones (2009). The sequence–structure relationship and protein function prediction. *Current Opinion in Structural Biology 19*, 357–362. doi:10.1016/j.sbi.2009.03.008.

Sarkar, S. (2008). Genomics, proteomics, and beyond. In S. Sarkar and A. Plutynski (Eds.), *A Companion to the Philosophy of Biology*, pp. 58–73. Blackwell Publishing Ltd.

Shapiro, J. A. (2011). *Evolution: a view from the 21st century.* FT Press Science.

Wouters, A. G. (2003). Four notions of biological function. *Studies in History and Philosophy of Biological and Biomedical Sciences 34*(4), 633–668. doi:10.1016/j.shpsc.2003.09.006.

Wright, L. (1973). Functions. *The Philosophical Review 82*(2), 139–168. doi:10.2307/2183766.

Wright, P. E. and H. J. Dyson (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology 293*, 321–331. doi:10.1006/jmbi.1999.3110.

Zhou, M., J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. Sachs, and Y. Liu (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature 495*, 111–115. doi:10.1038/nature11833.

Is Peer Review a Good Idea?*

Remco Heesen[†‡] Liam Kofi Bright[§]

September 19, 2018

Abstract

Pre-publication peer review should be abolished. We consider the effects that such a change will have on the social structure of science, paying particular attention to the changed incentive structure and the likely effects on the behavior of individual scientists. We evaluate these changes from the perspective of epistemic consequentialism. We find that where the effects of abolishing pre-publication peer review can be evaluated with a reasonable level of confidence based on presently available evidence, they are either positive or neutral. We conclude that on present evidence abolishing peer review weakly dominates the status quo.

*Both authors contributed equally. Thanks to Justin Bruner, Adrian Currie, Cailin O'Connor, and Jan-Willem Romeijn for valuable comments. RH was supported by an Early Career Fellowship from the Leverhulme Trust and the Isaac Newton Trust. LKB was supported by NSF grant SES 1254291.

[†]Department of Philosophy, School of Humanities, University of Western Australia, Crawley, WA 6009, Australia. Email: remco.heesen@uwa.edu.au.

[‡]Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK.

[§]Department of Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: liamkbright@gmail.com.

1 Introduction

Peer review plays a central role in contemporary academic life. It sits at the critical juncture where scientific work is accepted for publication or rejected. This is particularly clear when the results of scientific work are communicated to non-scientists, e.g., by journalists. The question "Has this been peer reviewed?" is commonly asked, and a positive answer is frequently taken to be a necessary and sufficient condition for the results to be considered serious science.

Given these circumstances, one might expect peer review to be an important topic in the philosophy of science as well. Peer review should arguably play a more prominent role in the debate about demarcation criteria (what separates science from other human pursuits?), as it seems to be used in practice exactly to differentiate scientific knowledge from other claims to knowledge, at least by journalists. Yet as far as we know, social-procedural accounts of science, like the one found in Longino (1990), remain in the minority and usually do not place great emphasis on peer review in particular. Aside from this particular debate, there are normative questions about the proper epistemic role of peer review and more practical questions about the extent to which it manages to fulfill them, all of which should interest philosophers of science.

But there has been surprisingly little work on peer review by philosophers of science. Most of what exists has focused on the role of biases in peer review, see for example Saul (2013, §2.1), Lee et al. (2013), Jukola (2017), Katzav and Vaesen (2017), and Heesen (2018). We are not aware of any philosophical discussion of the strengths and weaknesses of peer review as such (the above examples presuppose its overall legitimacy by discussing its implementation). Some work along these lines does exist outside of philosophy, either in the form of opinion pieces (Gowers 2017) or occasionally full-length articles (Smith 2006). Such work tends to be vague about the normative standard against which peer review or its alternatives are to be

evaluated, something we aim to remedy in section 2.

Here we bring together the work of philosophers of science (especially social epistemologists of science) who have written about the strengths and weaknesses of various aspects of the social structure of science and empirical work about the effects of peer review. We argue that where philosophers of science have claimed the social structure of science works well, their arguments tend to rely on things other than peer review, and that where specific benefits have been claimed for peer review, empirical research has so far failed to bear these out. Comparing this to the downsides of peer review, most prominently the massive amount of time and resources tied up in it, we conclude that we might be better off abolishing peer review.

Some brief clarifications. Our target is pre-publication peer review, that is the review of a manuscript intended for publication, where publication is withheld until one or more editors deem the manuscript to have successfully passed peer review. We set aside other uses of peer review (e.g., of grant proposals or conference abstracts) and we explicitly leave room for post-publication peer review, where manuscripts are published before review. Because of this last point, some readers may think that our terminology ('abolishing pre-publication peer review') suggests a more dramatic change than what we actually advocate. We invite such readers to substitute in their preferred terminology. We should also clarify that we use 'science' in a broad sense to include the natural sciences, the social sciences, and the humanities.

The overall structure of our argument is as follows. We think there are a number of clear benefits to abolishing pre-publication peer review. In contrast, while various benefits of the existing system (downsides of abolishing peer review) have been suggested, we do not think there exist any that have clear empirical support. Insofar as empirical research exists, it is ambiguous in some cases, and speaks relatively clearly against the claimed benefit of the existing system in others. While we admit to a number of cases where the evidence is ambiguous or simply lacking (see especially section 5), we claim

that the present state of the evidence suggests that abolishing pre-publication peer review would lead to a Pareto improvement: each factor considered is either neutral or favors our proposal.

Our primary aim here is to evaluate the current system, but we believe that is only really possible by comparing it to an alternative. We are not claiming that the proposal we put forward is the best of all possible alternatives. It has been constructed to be a system which could constitute a Pareto improvement over the current system. Given that it has not actually been implemented yet, we cannot guarantee it would work as advertised or what empirical properties it would have. But in offering a relatively specific alternative, we hope to get people thinking about real change, which pointing out problems with the present system has so far failed to do.

Even despite this proviso, we realize that ours is a strong claim, and our proposal a large change to the social structure of science. It is therefore important to highlight that our central claim concerns the balance of presently available evidence. We are not further claiming that the matter is so conclusively settled as to render further research superfluous or wasteful. On the contrary, we think there are a number of points in our argument where the presently available evidence is severely limited, and we take the calls for further empirical research we make in those places to be just as important a part of the upshot of our paper as our positive proposal. We hope, therefore, that even a skeptical reader will read on; if not to be convinced of the need of abolishing pre-publication peer review, then at least to see where in our view their future research efforts should concentrate if they are to shore up pre-publication peer review's claims to good epistemic standing.

2 Setting the Stage

The purpose of peer review is usually construed in terms of quality control. For example, Katzav and Vaesen (2017, 6) write "The epistemic role of peer

review is assessing the quality of research", and this seems to be a common sentiment per, e.g., Eisenhart (2002, 241) and Jukola (2017, 125). But how well does peer review succeed in its purpose of quality control? The empirical evidence (reviewed below) is mixed at best. As one prominent critic puts it, "we have little evidence on the effectiveness of peer review, but we have considerable evidence on its defects" (Smith 2006, 179).

Peer review's limited effectiveness would perhaps not be a problem if it required little time and effort from scientists. But in fact the opposite is true. Going from a manuscript to a published paper involves many hours of reviewing work by the assigned peer reviewers and a significant time investment from the editor handling the submission. The editor and reviewers are all scientists themselves, so the epistemic opportunity cost of their reviewing work is significant: instead of reviewing, they could be doing more science.

Given these two facts—high (epistemic) costs and unclear benefits—we raise the question whether it might be better to abolish pre-publication peer review. In the following we provide our own survey and assessment of the evidence that bears on this question. Our conclusions are not sympathetic to peer review. However, we encourage any proponents of peer review to give their own assessment. We only ask that any benefits claimed for peer review are backed up by empirical research, and that they are epistemic benefits, i.e., we ask for empirical evidence that peer review makes for better science on science's own terms.

We take the status quo to be as follows. The vast majority of scientific work is shared through journal publications, and the vast majority of journals uses some form of pre-publication peer review. Ordinarily this means that an editor assigns one to three peers (scientists whose expertise intersects the topic of the submission), who provide a report and/or verdict on the submission's suitability for publication. The peer reviews feed into the final judgment: the submission is accepted or rejected with or without revisions.

Our proposal is to abolish pre-publication peer review. Scientists them-

selves will decide when their work is ready for sharing. When this happens, they publish their work online on something that looks like a preprint archive (although the term "preprint" would not be appropriate under our proposal). Authors can subsequently publish updated versions that reply to questions and comments from other scientists, which may have been provided publicly or privately. Most journals will probably cease to exist, but the business of those that continue will be to create curated collections of previously published articles. Their process for creating these collections will presumably still involve peer review, but now of the post-publication variety.

Our proposal is in line with how certain parts of mathematics and physics already work: uploading a paper to arXiv is considered publishing it for most purposes, with journal peer review and publication happening almost as an afterthought (Gowers 2017). Indeed, journal publication can function as something like a prize, accruing glory to the scientist who achieves it but doing little to actually help spread or diffuse the idea beyond calling attention to something that has already been made public elsewhere. We are not aware of any detailed comparative studies of the effects these changes have had in those fields, so we will not rest any significant part of our argument on this case. But for those who worry that science will immediately and irrevocably fall apart without peer review, we point out that this does not appear to have happened in the relevant parts of mathematics and physics.

In the remainder of this paper we break down the consequences of our proposal. Our strategy here is to focus on a large number (hopefully all) aspects of the social structure of science that will be affected. In particular, the reader may already have a particular objection against our proposal in mind. We encourage such a reader to skip ahead to the section where this objection is discussed before reading the rest of the paper.

For example, one reader may think that peer review as currently practiced is important because it forces scientists to read and review each other's work, and without peer review they will spend less time on such tasks. This

is discussed in section 3.2. Another reader may worry that without peer review and the journal publications that go with them it will be more difficult to evaluate scientists for hiring or promotion (section 3.5). Yet another reader may be concerned about losing peer review's ability to prevent work of little merit from being published, or at least to sort papers into journals by epistemic merit so scientists can easily find good work (section 4.1). A fourth reader might think peer review plays an important role in detecting fraud or other scientific malpractice (section 4.2). A fifth reader may think the guarantee provided to outsiders when something has been peer reviewed is an important reason to preserve the status quo (section 5.1). And a sixth reader may want to point out that anonymized peer review gives relatively unknown scientists a chance at an audience by publishing in a prestigious journal, whereas on our proposal perhaps only antecedently prominent scientists will have their work read and engaged with (section 5.2).

Other aspects of the social structure of science that will be considered: whether and when scientists share their work (section 3.1), how many papers are published by women or men (section 3.3), library resources (section 3.4), the power of editors as gatekeepers (section 3.6), science's susceptibility to fads and fashions (section 4.3), and ways to get credit for scientific work other than through journal publications (section 4.4). In each case we evaluate whether the net effects of our proposal on that aspect can be expected to be positive. To tip our hand: aspects where we will claim a benefit are gathered in section 3, aspects where we expect little or no change are in section 4, and aspects that we consider neutral due to a present lack of evidence are in section 5.

In making these evaluations, we commit to a kind of epistemic consequentialism (cf. Goldman 1999). One may think of what we are doing as roughly analogous to the utilitarian principle, where for each issue our yardstick is whether pre-publication peer review shall generate the greatest amount of knowledge produced in the least amount of time. More specifically, we con-

sider changes in the incentive structure and expected behaviour of scientists, as well as other changes that would result from abolishing pre-publication peer review. We evaluate these changes in terms of their expected effect on the ability of the scientific community to produce scientific knowledge in an efficient manner. Working out in detail what such an epistemic consequentialism would look like would be very complicated, and we do not attempt the task here. For most of the issues we consider, we think that the calculus is sufficiently clear that fine details do not matter. Where it is unclear (the issues discussed in section 5) we think this results from ignorance of empirical facts about the likely effect of policies, rather than conceptual unclarity in the evaluative metric. So we do not need to use our consequentialist yardstick to settle any difficult tradeoffs. All we need for our purposes is to make it clear that we are evaluating the peer review system by how well it does in incentivizing efficient knowledge production.

What do we mean by the incentive structure of science, mentioned in the previous paragraph? This addresses the motivations of scientists. Scientists are rewarded for their contributions with credit, i.e., with recognition from their peers as expressed through such things as awards, citations, and prestigious publications (Merton 1957, Hull 1988, Zollman 2018). Scientific careers are largely built on the reputations scientists acquire in this way (Latour and Woolgar 1986, chapter 5). As a result, scientists engage in behaviors that improve their chances of credit (Merton 1969, Dasgupta and David 1994, Zollman 2018).

While individual scientists may be motivated by credit to different degrees (curiosity, the thrill of discovery, and philanthropic goals are important motivations for many as well), the effect on careers means that credit-maximizing behavior is to some extent selected for. Thus we think it important to ensure that our proposal does not negatively affect the incentives currently in place for scientists to work effectively and efficiently.

3 Benefits of Abolishing Peer Review

3.1 Sharing Scientific Results

An important feature of (academic) science is that there is a norm of sharing one's findings with the scientific community. This has been referred to as the communist norm (Merton 1942). In recent surveys, scientists by and large confirm both the normative force of the communist norm and their actual compliance (Louis et al. 2002, Macfarlane and Cheng 2008, Anderson et al. 2010). This norm is epistemically beneficial to the scientific community, as it prevents scientists from needlessly duplicating each other's work.

Will abolishing peer review affect this practice? In order to answer this question, we need to know what motivates scientists to comply with the communist norm, that is to share their work. On the one hand there is the feeling that they ought to share generated by the existence of the norm itself. There is no reason to expect this to be changed by abolishing peer review.

On the other hand there is the motivation generated by the desire for credit. According to the priority rule, the first scientist to publish a particular discovery gets the credit for it (Merton 1957, Dasgupta and David 1994, Strevens 2003). So a scientist who wants to get credit for her discoveries has an incentive to publish them as quickly as possible, in order to maximize her chances of being first. Recent work suggests that this applies even in the case of smaller, intermediate discoveries (Boyer 2014, Strevens 2017, Heesen 2017b). All of this helps motivate scientists to share their work.

If peer review were to be abolished, the communist norm and the priority rule would still be in effect, so scientists would still be motivated to share their work as quickly as possible. However, the following change would occur.

In the absence of pre-publication peer review, scientists would be able to share their discoveries more quickly. In the current system, peer review can hold up publication for significant amounts of time, especially in the case of fields with high rejection rates or long turnaround times. During this time,

other scientists cannot build on the work and may spend their time needlessly duplicating the work. Cutting out this lag by letting scientists publish their own work when they think it is ready will speed up scientific progress. While being faster is not always better (it may increase the risk of error, cf. Heesen 2017c), in this case delays in publication are reduced without any reduction in the time spent on the scientific work itself.

To some extent this already occurs. Scientists, especially well-connected scientists, already share preprints that make the community aware of their work in advance of publication. For people who regularly do this, practically speaking little would change upon adopting the system we advocate. However, our proposal turns pre-journal-publication dispersal of work from a privilege of a well-connected few into the norm for everyone.

On this point, then, abolishing peer review is a net positive, as scientists will still be incentivized to share their work as soon as possible, but the delays associated with pre-publication peer review are removed.

3.2 Time Allocation

The current system restricts the way scientists are allowed to spend their time. For each paper submitted to a journal, a number of scientists are conscripted into reviewing it, and at least one editor has to spend time on that paper as well.

On our proposal, scientists would be free to choose how much of their time to spend reading and reviewing others' work as compared to other scientific activities. Some scientists would spend less time reviewing, some scientists would spend more, and some would spend exactly as much as under the current system.

For scientists in the latter category our proposal makes no difference, while for scientists in the other two categories our proposal represents a net improvement of how they spend their time, at least in their own judgments. We think people are the best judges of how to use their own time and labor.

We thus trust scientists' decisions in these regards, and welcome changes that would render many scientists' choices about how to allocate their own labor independent of the preferences of the relatively small number of editorial gatekeepers.

So we assume that scientists are well-placed to judge how best to use their own abilities to meet the community's epistemic needs. We claim, moreover, that the reward structure of science is set up so as to make it in their interest to do so: the credit economy incentivizes scientists to spend their time on pursuits the epistemic value of which will be recognized by the community (Zollman 2018). Hence freeing up the way scientists allocate their time leads to net epistemic benefits to the scientific community.

One might object that journals perform a useful epistemic sorting role, telling scientists what is worth spending their time on. We will address these concerns in section 4.1.

One might think that this would lead scientists to spend significantly less time reading and reviewing others' work. If this is right, we still think it would be an overall improvement for the reasons mentioned above. But we also want to point out that this is not as obvious a consequence as it may seem. Here are two reasons to expect scientists to spend as much time or more reading and reviewing on our proposal. First, for many scientists reading and reviewing are intrinsically valuable and can help their own research. Second, the current system provides no particular incentive to read and review either: scientists agree to review only because they independently want to or because they feel an obligation to the research community. While no one scientist is conscripted, at the group level editors are going to keep going until they find someone. This can amount to picking whomever is most weak-willed or under some extra-epistemic social pressure. It is not obvious that this way of deciding who does the reviewing has much to recommend it. Any rewards that exist for reviewing will still exist on our proposal, and may be amplified by the possibility of making post-publication reviews public.

3.3 Gender Skew in Publications

Male scientists publish more, on average, than female scientists, a phenomenon known as the productivity puzzle or productivity gap (Zuckerman and Cole 1975, Valian 1999, Prpić 2002, Etzkowitz et al. 2008). Several explanations have been suggested, none of which are entirely satisfactory (see especially Etzkowitz et al. 2008, 409–412). Two of these explanations that are relevant to our concerns here are the direct effects of gender bias and the indirect effects of the expectation of gender bias.

There is some evidence of gender bias in peer review, although this is not unambiguous (see Lee et al. 2013, 7–8, Lee 2016, and references therein). Insofar as there is gender bias—in the sense of women's work being judged more negatively by peer reviewers—abolishing peer review will remove this and help level the playing field for men and women. We expect positive epistemic consequences from the removal of these arbitrarily different standards.

While the evidence of gender bias in peer review is not entirely clear-cut, there is good evidence that women *expect* to face gender bias in peer review (see Lee 2016, Bright 2017b, Hengel 2018, and references therein). In an effort to overcome this perceived bias, women tend to hold their own work to higher standards. Hengel (2018) provides evidence that women spend more time correcting stylistic aspects of their paper during peer review, presumably due to higher expectations of scrutiny on such apparently superficial elements of their work. On the plausible assumption that if women have higher standards for each paper they will produce fewer papers overall, this means that the mere expectation of gender bias can contribute to the productivity gap.

After abolishing peer review both women and men will hold their work primarily to their own individual standards of quality, and secondarily to their expectations of the response of the entire scientific community, but not to their expectations of the opinion of a small arbitrary group of gatekeepers. We do not know whether this will lead the women to behave more like the men (producing more papers) or the men to behave more like the women

(holding individual papers to a higher standard of quality). However, in line with our view above that scientists are well-placed to judge how best to spend their own time, we take it that any resulting change in behavior will be a net epistemic positive.

3.4 Library Resources

Journal subscription fees currently take up a large amount of library resources (RIN 2008, Van Noorden 2013). To summarize some key figures from the 2008 report: research libraries in the UK spent between £208,000 and £1,386,000 on journal subscriptions annually (and that was a decade ago, with subscriptions having risen substantially since). The cost for publishing and distributing a paper was estimated to be about £4,000, or about £6.4 billion per year in total. Savings from moving to author-paid open access were estimated at £561 million, about half of which would directly benefit libraries.

On our proposal, this is replaced by the cost of maintaining one or more online archives of scientific publications. The example of existing large preprint archives like arXiv and bioRxiv suggests that maintaining such archives can be done at a fraction of the cost currently spent on journal subscription fees. As a rough guideline, Van Noorden (2013) estimates maintenance costs of arXiv at just $10 per article. So our proposal involves significant savings on library resources, which could be used to expand collections, retain more or better trained staff, or other purposes that would be of epistemic benefit to the scientific community.

Two additional effects should be considered in relation to this. First, the fact that the online archive will be open access means that scientific publications will be available to everyone, not just to those with a library subscription or some other form of access to for-profit scientific journals.

Second, the fact that any value added by for-profit journals would be taken away. The two tasks currently carried out by journals that could

plausibly be supposed to add value to scientific publications are peer review and copy-editing (Van Noorden 2013). It is the purpose of all other sections of this paper to argue that peer review does not in fact (provably) add value, so we set that aside. This leaves copy-editing. We propose that libraries use some of the funds freed up from journal subscriptions to employ some copy-editors. Each university library would make copy-editing services available to the scientists employed at that university. We contend that, after paying for the maintenance of an online archive and a team of copy-editors, under our proposal libraries would still end up with more resources for other pursuits than under the current system.

We note that this particular advantage of our proposal is a bit more historically contingent than the others. There seems to be no particular reason why pre-publication peer review has to be implemented through for-profit journals, and if the open access movement has its way we might be able to free up these library resources without abolishing pre-publication peer review. But our proposal also achieves this goal, and so we count it as an advantage relative to the system as it is currently actually implemented.

3.5 Scientific Careers

The 'publish or perish' culture in science has been widely noted (e.g., Fanelli 2010). Universities judge the research productivity of scientists through their publications in (peer reviewed) journals, with some focusing more on 'quantity' (counting publications) and others on 'quality' (publishing in prestigious journals). Scientific journals and the system of pre-publication peer review thus play an important role in shaping scientific careers. What will become of this if peer review is abolished?

We note first that the 'publish or perish' culture is a subset of a larger system which we discussed above: the credit economy. Publishing in a journal is one way to receive credit for one's work, but there are others, most prominently citations and awards. Scientific careers depend on all of these,

with different institutions weighting quantity of publications, quality of publications, citation metrics, and awards and other honors differently.

Any of these types of credit represents some kind of recognition of the scholarly contributions of the scientist by her peers. But here we distinguish two types of credit, which we will call short-run credit and long-run credit. Getting a paper through peer review yields a certain amount of credit: more for more prestigious journals, less for less prestigious ones. But this is short-run credit in the following sense. The editor and the peer reviewers judge the technical adequacy and the potential impact of the paper, shortly after it is written. Their judgment is essentially a prediction of how much uptake the paper is likely to receive in the scientific community.

In contrast, citations (as well as awards, prizes, inclusion in anthologies or textbooks, etc.) represent long-run credit. They *are* the uptake the paper receives in the scientific community. Long-run credit is both a more considered opinion of the scientific importance of the paper and a more democratic one (citations can be made by anyone, and awards usually reflect a consensus in the scientific community, whereas peer review is normally done by up to three individuals). So long-run credit reflects a more direct and better estimate of the real epistemic value of a contribution to science.

So what would the effect of our proposal be? For better or worse, our proposal does not make it impossible for universities to use metrics to judge research productivity. While journal rankings and impact factors would disappear, citation metrics for individual scientists and papers would still be available. This may mean that universities stop judging their scientists based on the impact factors of the journals they publish in and start judging them on the actual citation impact of their papers. More generally, our proposal will decrease or remove the role of short-run credit in shaping career outcomes and increase the role of long-run credit, which we take to be a better measure of scientific importance. So we think this is an improvement on the status quo.

What about junior hires and related career decisions, where long-run credit may be absent or minimal? If abolishing peer review means completely getting rid of journals and the associated prestige rankings, this robs hiring departments of some information regarding the scientific importance of candidates' work. If this means those on the hiring side need to read and form an opinion of candidates' work for themselves, we do not think that is a bad thing. This would of course take time, but if journals and peer review are completely abolished, that just means the time spent reviewing the paper is transferred to the people considering hiring the scientist, which again, we do not think is a bad thing. In fact, since very few academics are on a hiring committee year after year, whereas referee requests are a constant feature while one is in the community, we think that even this added burden when hiring might still be a net time-saver for academics.

But it does not have to be that way. We never said journals and peer review have to be completely abolished—our proposal in section 2 explicitly suggests journal issues may still appear, but as curated collections of articles based on post-publication peer review. So short-run credit based on journal prestige need not disappear. It need not even be slower as there is no particular reason post-publication peer review needs to take longer than pre-publication peer review. But there is the added advantage that the paper is already published while it undergoes peer review, so the wider community outside the assigned reviewers also has a chance to respond before it is included in a journal.

3.6 The Power of Gatekeepers

The discussion immediately above touched on another effect, one that we think is worth bringing out as a benefit of our proposal in its own right. As mentioned our proposal suggests that in evaluating the importance of scientific work we decrease our reliance on short-run credit (journal prestige), with a corresponding increase in long-run credit (citations, among other things).

This means that the overall credit associated with a particular paper depends less on the judgments made by an editor and a small number of reviewers, and more on its actual uptake in the larger scientific community.

Editors in particular currently play a large role in determining which scientific work is worthy of attention, as they are a relatively small group of people with a deciding vote in the peer review process of a large number of papers. They are often referred to as gatekeepers for this reason (Crane 1967). Our proposal entails significantly decreasing both the prevalence and importance of this role. By replacing some of this importance with long-run credit, which comes from the scientific community as a whole, it makes the evaluation of scientific work a more democratic process. Not only is there some reason to think that democratic evaluation of scientific claims is more in line with general communal norms accepted within science (Bright et al. 2018), but general arguments from democratic theory and social epistemology of science give epistemic reason to welcome the increased independence of judgment and evaluation this would introduce (List and Goodin 2001, Heesen et al. forthcoming, Perović et al. 2016, 103–104).

4 Where Peer Review Makes No Difference

In this section we consider a number of aspects of the scientific incentive structure for which we think a case can be made that abolishing peer review will leave them basically unaffected. This serves partially to forestall objections to our proposal that we anticipate from defenders of the peer review system, and partially to avoid overstating our case—in some of what follows we argue that abolishing peer review will likely have no effect in cases where one might have expected it to be beneficial.

4.1 Epistemic Sorting

Given the stated purpose of peer review mentioned in section 2 the first and most apparent disadvantage of our proposal is that it would remove the epistemic filter on what enters into the scientific literature. One might worry that the scientific community would lose the ability to maintain its own epistemic standards, and thus the general quality of scientific research would be reduced. We argue here that despite the intuitive support this idea might have, the present state of the literature on scientific peer review does not support it.

Separate out two kind of epistemic standards one may hope that the peer review system maintains. First, that peer review allows us to identify especially meritorious work and place it in high profile journals, while ensuring that especially shoddy work is kept from being published. Call this the 'epistemic sorting' function of peer review. Second, that peer review allows for the early detection of fraudulent work or work that otherwise involves research misconduct. Call this the 'malpractice detection' function of peer review. We deal with each of these in turn.

Let us step back and ask why, from the point of view of epistemic consequentialism, one would want peer review to do any sort of epistemic sorting. We take the answer to be that epistemic sorting helps scientists fruitfully direct their time and energy by selecting the best work and bringing it to scientists' attention through publication in journals. They read and respond to that which is most likely to help them advance knowledge in their field.

How could peer review achieve this? One might hope that peer review functions by keeping bad manuscripts out of the published literature and letting good manuscripts in. This, however, is a non-starter. There are far too many journals publishing far too many things, with standards of publication varying far too wildly between them, for the sheer fact of having passed peer review somewhere to be all that informative as to the quality of a manuscript.

Instead, if peer review is to serve anything like this purpose it must be because reviewers are able (even if imperfectly) to discern the relative degree of scientific merit of a work, and sort it into an appropriately prestigious journal. Epistemic sorting happens not via the binary act of granting or withholding publication, but rather through sorting manuscripts into journals located on a prestige hierarchy that tracks scientific merit.

A necessary condition for epistemic sorting to work as advertised is that reviewers be reliable guides to the merit of the scientific work they review. Our first critique is that this necessary condition does not seem to be met. Investigation into reviewing practices has not generally found much inter-reviewer reliability in their evaluations (Peters and Ceci 1982, Ernst et al. 1993, Lee et al. 2013, 5–6). What this means is that one generally cannot predict what one reviewer will think of a manuscript by seeing what another reviewer thought. If there was some underlying epistemic merit scientists were accurately (even if falteringly) discerning by means of their reviews, one would expect there to be correlations in reviewers evaluations. However, this is not what we find. Indeed, one study of a top medical journal even found that "reviewers... agreed on the disposition of manuscripts at a rate barely exceeding what would be expected by chance" (Kravitz et al. 2010, 3). Findings like these are typical in the literature that looks at inter-reviewer reliability (for a review of the literature see Bornmann 2011, 207). The available evidence does not provide much support for the idea that pre-publication peer review detects the presence of some underlying quality.

Our second critique of the epistemic sorting idea speaks more directly to the ideal it tracks. We are not persuaded that the best way to direct scientists' attention is to continually alert them to the best pieces of individual work, and have them proportion their attention according to position on a prestige hierarchy. We take it the intuition behind this is a broadly meritocratic one. This intuition has been challenged by some modeling work (Zollman 2009). While Zollman retained some role for peer review, his model

still found that striving to select the best work for publication is not necessarily best from the perspective of an epistemic community; his model favored a greater degree of randomization.

We do not wish to rest our case on the results of one model which in any case does not fully align with our argument, but it highlights that the ideal of meritocracy stands in need of more defense than it is typically given. We take it that scientists most fruitfully direct their attention to that package of previous work and results which, when combined, provides them with the sort of information and perspectives they need to best advance their own epistemically valuable projects. It is a presently undefended assumption that this package of work should be composed of works which are themselves individually the most meritorious work, or that paying attention to the prestige hierarchy of journals and proportioning one's attention accordingly will be useful in constructing such a package. Hence, even if it did turn out that the peer review system could sort according to scientific merit, it is an underappreciated but important fact that this is not the end of the argument. Further defense of the purpose of this kind of epistemic sorting is needed from the point of view of epistemic consequentialism.

Before moving on we note a potential objection. Even if one did not think that peer review was detecting some underlying quality or interestingness, one might think that the process of feedback and revision which forms part of the peer review system would be beneficial to the epistemic quality of the scientific literature. In this way epistemic sorting may have a positive epistemic effect even if it fails in its primary task.

However, this returns us to the points regarding gatekeepers and time allocation from section 3. We are not opposed to scientists reading each other's work, offering feedback, and updating their work in light of that. This can indeed lead to improvements (Bornmann 2011, 203), though in this context it is worth noting the results of an experiment in the biomedical sciences, which found that attempting to attach the allure of greater prestige

to more epistemically high caliber work did little to actually improve the quality of published literature (Lee 2013). Fully interpreting these results would require discussion of the measures of quality used in such literature. We do not intend to do that here, since we do not intend to dispute the point that it is desirable for scientists to give feedback and respond to it.

We would expect this sort of peer-to-peer feedback to continue under a system without pre-publication peer review. Curiosity, informal networking, collegial responsibilities, and the credit incentives to engage with others' work and make use of new knowledge before others do; these would all be retained even without pre-publication peer review. What would be eliminated is the assignment of reviewing duties to papers that scientists did not independently decide were worth their time and attention, and the necessity of giving uptake to criticism (in order to publish) independently of an author's own assessment of the value of that feedback.

We thus conclude that, from the point of view of epistemic consequentialism, there is presently little reason to believe that a loss of the epistemic sorting function of pre-publication peer review would be a loss to science. Inclusion in the literature does not do much to vouch for the quality of a paper; the evidence does not favor the hypothesis that reviewers are selecting for some latent epistemic quality in order to sort into appropriate journals; and the ideal underlying the claimed benefits of epistemic sorting is dubious. While peer reviewers do give potentially valuable feedback, there is no particular reason to think that changes in how scientists decide to spend their time would make things worse in this regard, and (per our arguments in section 3) some reason to think that they would make things better.

4.2 Malpractice Detection

The other way peer review might uphold epistemic standards is through malpractice detection. However, once again, the literature does not support this. A number of prominent cases of fraudulent research managed to sail

through peer review. Upon investigation into the behavior of those involved it was found there was no reason to think that peer reviewers or editors were especially negligent in their duties (Grant 2002, 3). Peer reviewers report unwillingness to challenge something as fraudulent even where they have some suspicion that this is so, and avoid the charge (Francis 1989, 11–12). A criminologist who looked into fraudulent behavior in science reported that "virtually no fraudulent procedures have been detected by referees because reading a paper is neither a replication nor a lie-detecting device" (Ben-Yehuda 1986, 6). A more recent survey of the evidence found, at the least, no consistent pattern in journals' self-reported ability to detect and weed out fraudulent results (Anderson et al. 2013, 235).

Even if the prospect of peer review puts some people off committing fraud, the fact that it is so unreliable at detecting fraud suggests that this is a very fragile deterrence system indeed. Even this psychological deterrence would be rapidly undermined by more adventurous souls, or those pushed by desperation, since many would quickly learn that pre-publication peer review is a paper tiger.

Conversely, there are various ways for malpractice detection to operate in the absence of peer review. These include motive modification (Nosek et al. 2012, Bright 2017a), encouraging post-publication replication and scrutiny (Bruner 2013, Romero 2017), and the sterner inculcation of the norms of science coupled with greater expectation of oversight among coworkers (Braxton 1990). All of these methods of deterring fraud or meliorating its effects would still be available under our proposal.

What evidence we now have gives little reason to suppose that abolishing pre-publication peer review is any great loss to malpractice detection. Thus in this regard our proposal would make no great difference to the epistemic health of science. Combining this with the discussion of epistemic sorting, we conclude there is presently no reason to believe pre-publication peer review is adding much value to science by upholding epistemic standards.

4.3 Herding Behavior

Where above we argued that pre-publication peer review is not making a positive difference often claimed for it, in this section we downplay a potential benefit of our proposal. A consistent worry about scientific behavior is that it is subject to fads or, in any case, some sort of undesirable herding behavior (see, e.g. Chargaff 1976, Abrahamson 2009, Strevens 2013). A natural thought is that pre-publication peer review encourages this, since by its nature it means that to get new ideas out there one must convince one's peers that the work is impressive and interesting. It has thus been claimed that pre-publication peer review encourages unambitious within-paradigm work that unduly limits the range of scientific activity (Francis 1989, 12). Reducing the incentive to herd might thus be claimed as a potential benefit of our proposal. However, we are not convinced that it is pre-publication peer review that is doing the harmful work here.

As mentioned above, our proposal eliminates or significantly reduces the importance of short-run credit, the credit that accrues to one in virtue of publishing in a (more or less prestigious) scientific journal. Long-run credit, on the other hand, is left untouched. Under any sort of credit system, a scientist needs to do work that the community will pay attention to, build upon, and recognize her for. The mere fact that (she believes that) her peers are interested in a topic and liable to respond to it is thus still positive reason to adopt a topic. This is true even if the scientist would not judge that topic to be the best use of her time if she were (hypothetically) free from the social pressures and constraints of the scientific credit system.

The best that could be said about our proposal in this regard is that scientists would not specifically have to pass a jury of peers before getting their work out there. But given that we anticipate continued competition for the attention of scientific coworkers, it is hard to say what the net effect in encouraging more experimental or less conformist scientific work would be.

Whatever conformist effects the credit incentive has (see also the discus-

sion immediately below) do not depend on whether it is short- or long-run credit one seeks. The conformism comes from the fact that credit incentives focus scientists' attention on the predicted reaction of their fellow scientists to their work. Pre-publication peer review might make this fact especially salient by bringing manuscripts before a jury of peers before they may be entered into the literature. But even without pre-publication peer review the credit-seeking scientist must be focused on her peers' opinions. So there is no particular reason to think that removing the pre-publication scrutiny of manuscripts will free scientists from their own anticipations of the fads and fashions of their day.

4.4 Long-Run Credit

We end this section by noting that many of the effects of the credit economy of science studied by social epistemologists really concern long-run credit rather than the short-run credit affected by retaining or eliminating pre-publication peer review. This point is not restricted to herding behavior.

For instance, social epistemologists have studied both the incentive to collaborate, and various iniquities that can arise when scientists do not start with equal power when deciding who shall do what work and how they shall be credited (Harding 1995, Boyer-Kassem and Imbert 2015, Bruner and O'Connor 2017, O'Connor and Bruner forthcoming). Whether or not manuscripts would have to pass pre-publication peer review in order to enter the scientific literature, there would still be benefits in the long run to collaboration, and (alas) there would still be social inequalities that allow for iniquities to manifest in the scientific prestige hierarchy.

For another example, social epistemologists have studied the ways in which the credit incentive encourages different strategies for developing a research profile or molding one's scientific personality to be more or less risk-taking (Weisberg and Muldoon 2009, Alexander et al. 2015, Thoma 2015). Once again, pre-publication peer review plays no particular role in the analy-

sis. The incentives to differentiate oneself from one's peers (without straying too far from the beaten path) and to mold one's personality accordingly exist independently of pre-publication peer review.

Two especially influential streams of work in the social epistemology of science have been the study of the division of cognitive labor (Kitcher 1990, Strevens 2003), and the role of credit in providing a spur to work in situations with a risk of under-production (Dasgupta and David 1994, Stephan 1996). These two streams have directed the focus of the field, and have formed some of the chief defenses of the credit economy of science as it now stands (but see Zollman 2018, for a more critical take).

We mention them here because pre-publication peer review or short-run credit again plays no particular role in the analyses offered by these papers. What drives their results is scientists' expectation that genuine scientific achievement will be recognized with credit. As we have argued above, it is long-run credit that best tracks genuine scientific achievement, and so it is long-run rather than short-run credit that grounds scientists' expectation in this regard. So in social epistemologists' most prominent defenses of the credit economy of science, long-run credit (while not named such) is the mechanism underlying the claimed epistemic benefits of the credit economy.

5 Difficulties For Our Proposal

We have discussed some benefits that would predictably accrue from abolishing peer review and some ways in which its apparent benefits are either under-evidenced or better attributed to the effects of long-run credit, which our proposal leaves untouched. We now discuss some cases which we take to be more problematic for our proposal—but by this point we hope to have at least convinced the reader that pre-publication peer review rests on shakier theoretical grounds than its widespread acceptance may lead one to suppose.

5.1 A Guarantee For Outsiders

One purpose pre-publication peer review serves is providing a guarantee to interested but non-expert parties. Science journalists, policy makers, scientists from outside the field the manuscript is aimed at, or interested non-scientists can take the fact that something has passed peer review as a stamp of approval from the field. At a minimum, peer review guarantees that outsiders are focusing on work that has convinced at least one relatively disinterested expert that the manuscript is worthy of public viewing. Given that there are real dangers to irresponsible science journalism or public action that is seen to be based on science that is not itself trustworthy (Bright 2018, §4), and that it is hard for non-experts to make the relevant judgment calls themselves, having a social mechanism to provide this kind of guarantee for outsiders is useful.

It is difficult to predict in advance what norms would come to exist for science journalists in the absence of pre-publication peer review. We thus first and foremost call for empirical research on this issue, possibly by studying what has happened in parts of mathematics and physics that already operate broadly along the lines we suggest (Gowers 2017).

However, against the presumption that things would be worse, we have two points to make. As the recent replication crisis has made clear, the value of peer review as a stamp of approval should not be overstated. There are reasons to doubt that peer review reliably succeeds in filtering out false results. We give three of them. First, peer reviewers face difficulties in actually assessing manuscripts—and just about anything can pass peer review eventually—as discussed under the heading of 'epistemic sorting' in section 4.1. Second, there are problems with the standards we presently use to evaluate manuscripts, in particular with the infamous threshold for statistical significance used in many fields (Ioannidis 2005, Benjamin et al. 2018). And third, deeper features of the incentive structure of science make replicability problems endemic (Smaldino and McElreath 2016, Heesen 2017c). Using

peer review as a stamp of approval may just be generating expert overconfidence (Angner 2006), without the epistemic benefits of greater reliability that would back this confidence up.

For the second part of our reply, recall that it is only pre-publication peer review that we seek to eliminate. We do not object to post-publication peer review resulting in papers being selected for inclusion in journals which mark the community's approval of such work, ideally after due and broad-based evaluation. If some such system were implemented then outsiders could use inclusion in such a journal as their marker of whether work is soundly grounded in the relevant science.

If such a stamp of approval from a journal or other communally recognized institution only comes a number of months or years after something is first published then we would expect it to represent a more well-considered judgment. Note that this would not necessarily slow the diffusion of knowledge as under the present system the same paper would have spent time hidden from view going through pre-publication peer review. The end result might not even be all that different from what happens in the present system, except that post-publication peer review would take into account more of the response or uptake from the wider scientific community. Thus it would more closely approximate the considered judgment of the community, as ultimately reflected in the long-run credit accorded to the paper.

5.2 A Runaway Matthew Effect

The second problem we are less confident we can deal with is that of exacerbating the Matthew effect. This is the phenomenon, first identified by Merton (1968), of antecedently more famous authors being credited more for work done simultaneously or collaboratively, even if the relative size or skill of their contribution does not warrant a larger share of the reward. Arguably the present system helps put a damper on the Matthew effect, allowing a junior or less prestigious author to secure attention for their work by publishing

in a high profile journal. Without such a mechanism to grab the attention of the field, perhaps scientists would just decide what to pay attention to based on their prior knowledge of the author or recommendation from others. This would strengthen the effects of networks of patronage and prestige bias favoring fancy universities. Thus squandering valuable opportunities to learn from those who were not initially lucky in securing a prestigious position or patronage from the already established.

While some have defended the Matthew effect (Strevens 2006), we will not go that route in defending our proposal for two reasons. First, the Matthew effect can perpetuate iniquities that themselves harm the generation and dissemination of knowledge (Bruner and O'Connor 2017). Second, even if it could be justified at the level of individual publications, its long-term effects are epistemically harmful. The scientific community allocates the resources necessary for future work on the basis of its recognition of past performance. So if there is excess reward for some and unfair passing over of others at the present stage of inquiry, this will ramify through to future rounds of inquiry, misallocating resources to people whose accomplishments do not fully justify their renown (Heesen 2017a). Hence on grounds of epistemic consequentialism we take seriously the problem of a runaway Matthew effect.

As mentioned, due to the pressures of credit-seeking and their own curiosity, scientists would still have incentive to read others' work and adapt it to suit their own projects. There is always a chance that valuable knowledge may be gathered from the work of one who has been ignored, which could provide an innovative edge. To some extent this creates opportunities for arbitrage: if the Matthew effect ever became especially severe there would be a credit incentive to specialize in seeking out the work of scientists who are not getting much attention. The lesson here is that the Matthew effect can only ever be so severe, before the credit incentive starts providing counter-veiling motivations.

However, this does not fully solve our problem. Moreover, so long as

resource allocation is tied to recognition of past performance the differences in recognition generated by the Matthew effect can and often do become self-fulfilling prophesies, as those with more gain the resources to do better in the future, and those without are starved of the resources necessary to show their worth.

It is not clear where to go from here. From the above it may seem like a solution would be to pair our proposal with a call to loosen the connection between recognition of a scientist's greatness based on their past performance and resource allocation. Indeed, this may well be independently motivated (Avin forthcoming, Heesen 2017a, §6). However, even short of this far-reaching change, we feel at present that this matter deserves more study rather than any definitive course of action.

Our present thought is that this is a very speculative objection, and there is no empirical evidence to back up the claim that eliminating pre-publication peer review will have dire consequences in this regard. In particular, while the present system may (rarely) allow a relative outsider to make a big splash, the common accusation of prestige bias in peer review (Lee et al. 2013, 7) suggests that on the whole pre-publication peer review may contribute to the Matthew effect rather than curtailing it.

More specifically, the Matthew effect can be made worse by peer review when anonymity breaks down in ways that systematically favor antecedently famous scientists. If this gives famous scientists more opportunities to publish papers, then our system may provide welcome relief, since it allows more people to get their papers out there. Hence whether our proposal makes the Matthew effect worse or better depends on whether the stronger influence would be who gets into the conversation (for which pre-publication peer review can exacerbate the Matthew effect), or who gets listened to once the conversation has begun (for which our proposal looks more problematic). Presently we cannot say which is the more significant effect. So, while we grant that a runaway Matthew effect may occur under our system, we prefer

to stress that at this point it is just not known whether the Matthew effect will be worse with or without pre-publication peer review.

What we propose is a large change, involving freeing up a lot of time and opening it up to more self-direction on the part of scientists, and it is not clear what sort of institutional changes it would be paired with. With more study of epistemic mechanisms designed especially to promote the work of junior or less prestigious scientists there might be found some way of surmounting the problem of a runaway Matthew effect, should it arise. Ultimately, only empirical evidence can settle these questions. Given the clear benefits and the unclear downsides of our proposal, we hope at minimum to inspire a more experimental attitude towards peer review.

6 Conclusion

Pre-publication peer review is an enormous sink of scientists' time, effort, and resources. Adopting the perspective of epistemic consequentialism and reviewing the literature on the philosophy, sociology, and social epistemology of science, we have argued that we can be confident that there would be benefits from eliminating this system, but have no strong reasons to think there will be disadvantages. There is hence a kind of weak dominance or Pareto argument in favor of our proposal.

To simplify things, imagine forming a decision matrix, with rows corresponding to 'Keeping pre-publication peer review' and 'Eliminating pre-publication peer review'. The columns would each be labeled with an issue studied by science scholars which we have surveyed here: gender bias in the literature, speed of dissemination of knowledge, efficient allocation of scientists' time and attention, etc. For each column, if there is a clear reason to think that either keeping or eliminating pre-publication scientific peer review does better according to the standards of epistemic consequentialism, place a 1 in the row of that option, and a 0 in the other. If there is no reason to

favor either according to present evidence, put a 0 in both rows.

Our present argument could then be summarized with: as it stands, the only 1s in such a table would appear in the row for eliminating pre-publication peer review. We thus advocate eliminating pre-publication peer review. Journals could still exist as a forum for recognizing and promoting work that the community as a whole perceives as especially meritorious and wishes to recommend to outsiders. Scientists would still have every reason to read, respond to, and consider the work of their peers; pre-publication peer review is not the primary drive behind either the intellect's curiosity or the will's desire for recognition, and either of those suffice to motivate such behaviors.

The overall moral to be drawn mirrors that of our invocation of the importance of long-run over short-run credit. The best guarantor of the long run epistemic health of science is science: the organic engagement with each others' ideas and work that arises from scientists deciding for themselves how to allocate their cognitive labor, and doing the hard work of replicating and considering from new angles those ideas that have been opened up to the scrutiny of the community. All this would continue without pre-publication peer review, and the best you can say for the system that currently uses up so much of our time and resources is that it often fails to get in the way.

References

Eric Abrahamson. Necessary conditions for the study of fads and fashions in science. *Scandinavian Journal of Management*, 25(2):235–239, 2009. doi: 10.1016/j.scaman.2009.03.005. URL `http://dx.doi.org/10.1016/j.scaman.2009.03.005`.

Jason McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. Epistemic landscapes, optimal search, and the division of cognitive

labor. *Philosophy of Science*, 82(3):424–453, 2015. doi: 10.1086/681766. URL http://dx.doi.org/10.1086/681766.

Melissa S. Anderson, Emily A. Ronning, Raymond De Vries, and Brian C. Martinson. Extending the Mertonian norms: Scientists' subscription to norms of research. *The Journal of Higher Education*, 81(3):366–393, 2010. ISSN 1538-4640. doi: 10.1353/jhe.0.0095. URL https://muse.jhu.edu/journals/journal_of_higher_education/v081/81.3.anderson.html.

Melissa S. Anderson, Marta A. Shaw, Nicholas H. Steneck, Erin Konkle, and Takehito Kamata. Research integrity and misconduct in the academic profession. In Michael B. Paulsen, editor, *Higher Education: Handbook of Theory and Research*, volume 28, chapter 5, pages 217–261. Springer, Dordrecht, 2013. doi: 10.1007/978-94-007-5836-0_5. URL http://dx.doi.org/10.1007/978-94-007-5836-0_5.

Erik Angner. Economists as experts: Overconfidence in theory and practice. *Journal of Economic Methodology*, 13(1):1–24, 2006. doi: 10.1080/13501780600566271. URL http://dx.doi.org/10.1080/13501780600566271.

Shahar Avin. Centralised funding and epistemic exploration. *The British Journal for the Philosophy of Science*, forthcoming. doi: 10.1093/bjps/axx059. URL http://dx.doi.org/10.1093/bjps/axx059.

Nachman Ben-Yehuda. Deviance in science: Towards the criminology of science. *British Journal of Criminology*, 26(1):1–27, 1986. doi: 10.1093/oxfordjournals.bjc.a047577. URL http://dx.doi.org/10.1093/oxfordjournals.bjc.a047577.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical

significance. *Nature Human Behaviour*, 2(1):6–10, 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0189-z. URL http://dx.doi.org/10.1038/s41562-017-0189-z.

Lutz Bornmann. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245, 2011. ISSN 1550-8382. doi: 10.1002/aris.2011.1440450112. URL http://dx.doi.org/10.1002/aris.2011.1440450112.

Thomas Boyer. Is a bird in the hand worth two in the bush? Or, whether scientists should publish intermediate results. *Synthese*, 191(1): 17–35, 2014. ISSN 0039-7857. doi: 10.1007/s11229-012-0242-4. URL http://dx.doi.org/10.1007/s11229-012-0242-4.

Thomas Boyer-Kassem and Cyrille Imbert. Scientific collaboration: Do two heads need to be more than twice better than one? *Philosophy of Science*, 82(4):667–688, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/682940.

John M. Braxton. Deviance from the norms of science: A test of control theory. *Research in Higher Education*, 31(5):461–476, 1990. doi: 10.1007/BF00992713. URL http://dx.doi.org/10.1007/BF00992713.

Liam Kofi Bright. On fraud. *Philosophical Studies*, 174(2):291–310, 2017a. ISSN 1573-0883. doi: 10.1007/s11098-016-0682-7. URL http://dx.doi.org/10.1007/s11098-016-0682-7.

Liam Kofi Bright. Decision theoretic model of the productivity gap. *Erkenntnis*, 82(2):421–442, 2017b. ISSN 1572-8420. doi: 10.1007/s10670-016-9826-6. URL http://dx.doi.org/10.1007/s10670-016-9826-6.

Liam Kofi Bright. Du Bois' democratic defence of the value free ideal. *Synthese*, 195(5):2227–2245, 2018. ISSN 1573-0964.

doi: 10.1007/s11229-017-1333-z. URL http://dx.doi.org/10.1007/s11229-017-1333-z.

Liam Kofi Bright, Haixin Dang, and Remco Heesen. A role for judgment aggregation in coauthoring scientific papers. *Erkenntnis*, 83(2):231–252, 2018. ISSN 1572-8420. doi: 10.1007/s10670-017-9887-1. URL http://dx.doi.org/10.1007/s10670-017-9887-1.

Justin Bruner and Cailin O'Connor. Power, bargaining, and collaboration. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 7, pages 135–157. Oxford University Press, Oxford, 2017.

Justin P. Bruner. Policing epistemic communities. *Episteme*, 10(4):403–416, Dec 2013. ISSN 1750-0117. doi: 10.1017/epi.2013.34. URL http://dx.doi.org/10.1017/epi.2013.34.

Erwin Chargaff. Triviality in science: A brief meditation on fashions. *Perspectives in Biology and Medicine*, 19(3):324–333, 1976. doi: 10.1353/pbm.1976.0011. URL http://dx.doi.org/10.1353/pbm.1976.0011.

Diana Crane. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 2(4):195–201, 1967. ISSN 00031232. URL http://www.jstor.org/stable/27701277.

Partha Dasgupta and Paul A. David. Toward a new economics of science. *Research Policy*, 23(5):487–521, 1994. ISSN 0048-7333. doi: 10.1016/0048-7333(94)01002-1. URL http://www.sciencedirect.com/science/article/pii/0048733394010021.

Margaret Eisenhart. The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, 32(2):241–255, 2002. ISSN 1573-1898. doi: 10.1023/A:1016082229411. URL http://dx.doi.org/10.1023/A:1016082229411.

Edzard Ernst, T. Saradeth, and Karl Ludwig Resch. Drawbacks of peer review. *Nature*, 363(6427):296, 1993. doi: 10.1038/363296a0. URL http://dx.doi.org/10.1038/363296a0.

Henry Etzkowitz, Stefan Fuchs, Namrata Gupta, Carol Kemelgor, and Marina Ranga. The coming gender revolution in science. In Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, editors, *The Handbook of Science and Technology Studies*, chapter 17, pages 403–428. MIT Press, Cambridge, third edition, 2008. ISBN 9780262083645.

Daniele Fanelli. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, 5(4):e10271, Apr 2010. doi: 10.1371/journal.pone.0010271. URL http://dx.doi.org/10.1371/journal.pone.0010271.

Jere R. Francis. The credibility and legitimation of science: A loss of faith in the scientific narrative. *Accountability in Research: Policies and Quality Assurance*, 1(1):5–22, 1989. doi: 10.1080/08989628908573770. URL http://dx.doi.org/10.1080/08989628908573770.

Alvin I. Goldman. *Knowledge in a Social World*. Oxford University Press, Oxford, 1999. ISBN 0198237774.

Timothy Gowers. The end of an error? *The Times Literary Supplement*, October 2017. URL https://www.the-tls.co.uk/articles/public/the-end-of-an-error-peer-review/. Editorial.

Paul M. Grant. Scientific credit and credibility. *Nature Materials*, 1:139–141, 2002. doi: 10.1038/nmat756. URL http://dx.doi.org/10.1038/nmat756.

Sandra Harding. "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3):331–349, 1995. doi: 10.1007/BF01064504. URL http://dx.doi.org/10.1007/BF01064504.

Remco Heesen. Academic superstars: Competent or lucky? *Synthese*, 194 (11):4499–4518, 2017a. ISSN 1573-0964. doi: 10.1007/s11229-016-1146-5. URL http://dx.doi.org/10.1007/s11229-016-1146-5.

Remco Heesen. Communism and the incentive to share in science. *Philosophy of Science*, 84(4):698–716, 2017b. ISSN 0031-8248. doi: 10.1086/693875. URL http://dx.doi.org/10.1086/693875.

Remco Heesen. Why the reward structure of science makes reproducibility problems inevitable. Manuscript, September 2017c. URL http://remcoheesen.files.wordpress.com/2015/03/rewards-and-reproducibility2.pdf.

Remco Heesen. When journal editors play favorites. *Philosophical Studies*, 175(4):831–858, 2018. ISSN 0031-8116. doi: 10.1007/s11098-017-0895-4. URL http://dx.doi.org/10.1007/s11098-017-0895-4.

Remco Heesen, Liam Kofi Bright, and Andrew Zucker. Vindicating methodological triangulation. *Synthese*, forthcoming. ISSN 1573-0964. doi: 10.1007/s11229-016-1294-7. URL http://dx.doi.org/10.1007/s11229-016-1294-7.

Erin Hengel. Publishing while female: Are women held to higher standards? Evidence from peer review. Manuscript, August 2018. URL http://www.erinhengel.com/research/publishing_female.pdf.

David L. Hull. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. The University of Chicago Press, Chicago, 1988. ISBN 0226360504.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, Aug 2005. doi: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

Saana Jukola. A social epistemological inquiry into biases in journal peer review. *Perspectives on Science*, 25(1):124–148, 2017. doi: 10.1162/POSC_a_00237. URL http://dx.doi.org/10.1162/POSC_a_00237.

J. Katzav and K. Vaesen. Pluralism and peer review in philosophy. *Philosophers' Imprint*, 17(19):1–20, 2017. URL http://hdl.handle.net/2027/spo.3521354.0017.019.

Philip Kitcher. The division of cognitive labor. *The Journal of Philosophy*, 87(1):5–22, 1990. ISSN 0022362X. URL http://www.jstor.org/stable/2026796.

Richard L. Kravitz, Peter Franks, Mitchell D. Feldman, Martha Gerrity, Cindy Byrne, and William M. Tierney. Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? *PLoS ONE*, 5(4):e10072, 2010. doi: 10.1371/journal.pone.0010072. URL http://dx.doi.org/10.1371/journal.pone.0010072.

Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, second edition, 1986.

Carole J. Lee. The limited effectiveness of prestige as an intervention on the health of medical journal publications. *Episteme*, 10(4):387–402, 2013. doi: 10.1017/epi.2013.35. URL http://dx.doi.org/10.1017/epi.2013.35.

Carole J. Lee. Revisiting current causes of women's underrepresentation in science. In Jennifer Saul and Michael Brownstein, editors, *Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology*, chapter 2.5, pages 265–282. Oxford University Press, Oxford, 2016. doi: 10.1093/acprof:oso/9780198713241.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780198713241.001.0001.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

Christin List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001. ISSN 1467-9760. doi: 10.1111/1467-9760.00128. URL http://dx.doi.org/10.1111/1467-9760.00128.

Helen E. Longino. *Science as Social Knowledge*. Princeton University Press, 1990.

Karen Seashore Louis, Lisa M. Jones, and Eric G. Campbell. Macroscope: Sharing in science. *American Scientist*, 90(4):304–307, 2002. ISSN 00030996. URL http://www.jstor.org/stable/27857685.

Bruce Macfarlane and Ming Cheng. Communism, universalism and disinterestedness: Re-examining contemporary support among academics for Merton's scientific norms. *Journal of Academic Ethics*, 6(1):67–78, 2008. ISSN 1570-1727. doi: 10.1007/s10805-008-9055-y. URL http://dx.doi.org/10.1007/s10805-008-9055-y.

Robert K. Merton. A note on science and democracy. *Journal of Legal and Political Sociology*, 1(1–2):115–126, 1942. Reprinted in Merton (1973, chapter 13).

Robert K. Merton. Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22(6):635–659, 1957. ISSN 00031224. URL http://www.jstor.org/stable/2089193. Reprinted in Merton (1973, chapter 14).

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63,

1968. ISSN 00368075. URL http://www.jstor.org/stable/1723414. Reprinted in Merton (1973, chapter 20).

Robert K. Merton. Behavior patterns of scientists. *The American Scholar*, 38 (2):197–225, 1969. ISSN 00030937. URL http://www.jstor.org/stable/41209646. Reprinted in Merton (1973, chapter 15).

Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago, 1973. ISBN 0226520919.

Brian A. Nosek, Jeffrey R. Spies, and Matt Motyl. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012. doi: 10.1177/1745691612459058. URL http://pps.sagepub.com/cgi/content/abstract/7/6/615.

Cailin O'Connor and Justin Bruner. Dynamics and diversity in epistemic communities. *Erkenntnis*, forthcoming. ISSN 1572-8420. doi: 10.1007/s10670-017-9950-y. URL http://dx.doi.org/10.1007/s10670-017-9950-y.

Slobodan Perović, Sandro Radovanović, Vlasta Sikimić, and Andrea Berber. Optimal research team composition: data envelopment analysis of Fermilab experiments. *Scientometrics*, 108(1):83–111, 2016. doi: 10.1007/s11192-016-1947-9. URL http://dx.doi.org/10.1007/s11192-016-1947-9.

Douglas P. Peters and Stephen J. Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2):187–195, 1982. doi: 10.1017/S0140525X00011213. URL http://dx.doi.org/10.1017/S0140525X00011213.

Katarina Prpić. Gender and productivity differentials in science. *Scientometrics*, 55(1):27–58, 2002. ISSN 0138-9130. doi: 10.1023/A:1016046819457. URL http://dx.doi.org/10.1023/A:1016046819457.

RIN. Activities, costs and funding flows in the scholarly communications system in the UK. Technical report, Cambridge Economic Policy Associates on behalf of the Research Information Network, 2008. URL http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/activities-costs-and-funding-flows-scholarly-commu.

Felipe Romero. Novelty versus replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, 84(5):1031–1043, 2017. ISSN 0031-8248. doi: 10.1086/694005. URL http://dx.doi.org/10.1086/694005.

Jennifer Saul. Implicit bias, stereotype threat, and women in philosophy. In Katrina Hutchison and Fiona Jenkins, editors, *Women in Philosophy: What Needs to Change?*, chapter 2, pages 39–60. Oxford University Press, Oxford, 2013.

Paul E. Smaldino and Richard McElreath. The natural selection of bad science. *Royal Society Open Science*, 3(9), 2016. doi: 10.1098/rsos.160384. URL http://rsos.royalsocietypublishing.org/content/3/9/160384.

Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182, 2006. URL http://jrs.sagepub.com/content/99/4/178.short.

Paula E. Stephan. The economics of science. *Journal of Economic Literature*, 34(3):1199–1235, 1996. URL http://www.jstor.org/stable/2729500.

Michael Strevens. The role of the priority rule in science. *The Journal of Philosophy*, 100(2):55–79, 2003. ISSN 0022362X. URL http://www.jstor.org/stable/3655792.

Michael Strevens. The role of the Matthew effect in science. *Studies in History and Philosophy of Science Part A*, 37(2):159–170, 2006. ISSN 0039-3681. doi: http://dx.doi.org/10.1016/j.shpsa.2005.07.009. URL http://www.sciencedirect.com/science/article/pii/S0039368106000252.

Michael Strevens. Herding and the quest for credit. *Journal of Economic Methodology*, 20(1):19–34, 2013. doi: 10.1080/1350178X.2013.774849. URL http://dx.doi.org/10.1080/1350178X.2013.774849.

Michael Strevens. Scientific sharing: Communism and the social contract. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 1. Oxford University Press, Oxford, 2017. URL https://philpapers.org/rec/STRSSC-2.

Johanna Thoma. The epistemic division of labor revisited. *Philosophy of Science*, 82(3):454–472, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/681768.

Virginia Valian. *Why So Slow? The Advancement of Women*. MIT Press, Cambridge, 1999. ISBN 9780262720311.

Richard Van Noorden. The true cost of science publishing. *Nature*, 495 (7442):426–429, 2013. ISSN 0028-0836. doi: 10.1038/495426a. URL http://dx.doi.org/10.1038/495426a.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/644786.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6(2):185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://dx.doi.org/10.3366/E174236000900063X.

Kevin J. S. Zollman. The credit economy and the economic rationality of science. *The Journal of Philosophy*, 115(1):5–33, 2018. doi: 10.5840/jphil201811511. URL http://dx.doi.org/10.5840/jphil201811511.

Harriet Zuckerman and Jonathan R. Cole. Women in American science. *Minerva*, 13(1):82–102, 1975. ISSN 1573-1871. doi: 10.1007/BF01096243. URL http://dx.doi.org/10.1007/BF01096243.

Epistemic Loops and Measurement Realism

Alistair M. C. Isaac

Abstract

Recent philosophy of measurement has emphasized the existence of both diachronic and synchronic "loops," or feedback processes, in the epistemic achievements of measurement. A widespread response has been to conclude that measurement outcomes do not convey interest-independent facts about the world, and that only a coherentist epistemology of measurement is viable. In contrast, I argue that a form of measurement realism is consistent with these results. The insight is that antecedent structure in measuring spaces constrains our empirical procedures such that successful measurement conveys a limited, but veridical knowledge of "fixed points," or stable, interest-independent features of the world.

§1 Introduction

Recent philosophy of measurement has employed detailed case studies to highlight the complex, iterative process by which measurement practices are refined. Typically, these examples are taken to support some form of epistemic coherentism, on which the validation of measurement procedures, and thus their epistemic import, is irreducibly infected by the contingent history of their development in aid of human interests. This coherentism in turn undermines *measurement realism*, the view that outcomes of successful measurement practices veridically represent objective (i.e. interest-independent) features of the world. For instance, van Fraassen (2008) takes the historical contingency of measurement practice to support empiricism, and Chang (2012) argues that only a pragmatic, interest-relative "realism" about measurement outcomes is plausible, not one which interprets them as corresponding to objective features in the world. More generally, Tal (2013) identifies coherentism as a major trend within contemporary philosophy of measurement.

I argue that the iterative and coherentist features of measurement practice these authors rightly emphasize are nevertheless consistent with realism about measurement outcomes. Nevertheless, my position contrasts significantly with that of other measurement realists, such as Byerly and Lazara (1973) or Michell (2005), who take measurement realism to be continuous with global scientific realism. On their view, measurement realism is a *stronger* position than traditional realism, imputing reality not only to theoretical objects and laws, but also to their quantitative character. The view defended here reverses this priority, articulating a realism about measurement outcomes *weaker* than traditional realism. In particular, I argue that the convergent assignment of increasingly precise values that constitutes successful measurement serves as incontrovertible evidence for *fixed points* in the world — features or events standing in stable quantitative relationships — even though the evidence it provides for any non-numerical theoretical description of these points is defeasible. The insight here is that measurement is more evidentially demanding than traditional confirmation, i.e. it requires a greater contribution from the interest-independent world to succeed than mere qualitative experiments. I argue that this greater evidential demand is a consequence of the

antecedent numerical structure in which measurement outcomes are represented. This antecedent structure blocks the possibility of gerrymandered categories that crosscut the joints of nature. Consequently, successful measurement constitutes a substantive enough epistemic achievement that we may legitimately "factor out" the contribution to success made by human interests, and accept the outcome as representing an objective feature of the world.

After surveying the motivations for measurement coherentism, I elaborate on the notion of "successful" measurement, and why it poses a challenge to coherentism. The paper concludes with a more careful articulation of the distinctive features of fixed point realism.

§2 Epistemic Loops in Measurement Practice

Contemporary measurement coherentism is motivated by two types of case study, each identifying a different kind of epistemic "loop," or feedback process driving knowledge formation. Chang and van Fraassen emphasize diachronic examples of epistemic iteration, where the feedback process extends over several stages of mutual influence between theory change and refinement of measurement practice. A different kind of epistemic loop has been discussed by Tal and metrologist Mari, who highlight the role of models in the calibration of measurement instruments and the assignment of quantity values, illustrating a synchronic epistemic interdependence between theory and measurement.

§2.1 Epistemic Iteration

Chang (2004) defines *epistemic iteration* as "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals" (45). He takes this process to support a "progressive coherentism": on the one hand, the criteria for measurement success are internal to a practice, so scientific knowledge does not rest on an independent foundation; on the other hand, these internal criteria may be used to evaluate new practices as improvements or refinements on their predecessors, thereby allowing for scientific progress (in contrast to traditional coherentism, Chang 2007). In the context of measurement, this means that later measurement practices may be understood as in some sense "better" than earlier ones, yet these "epistemic achievements" should not be cashed out as greater degree of correspondence to quantities in the world.

For instance, thermometry as a practice begins with subjective assignments of relative heat on the basis of our bodily experiences. Noticing that fluids appear to change volume in rough correspondence with these subjective sensations, one may construct a thermoscope, or device allowing comparison of relative fluid volumes in different circumstances. Already a theoretical leap is required to identify the cause of these changes in relative volume with the cause of our differing subjective sensations, especially given the discrepancies between these sensations and our thermoscopic readings (e.g. contrary to experience, caves are warmer in summer than they are in winter). Nevertheless, the move to the thermoscope constitutes an epistemic achievement, in the sense that it allows for greater regularity in the assignment of relative temperatures, both

across contexts and across observers. A similar pattern is seen in the move from thermoscope to thermometer, which enables assignment of numbers to temperatures. Numerical representation constitutes a yet greater epistemic achievement, insofar as it allows comparison of temperature assignments across devices. Nevertheless, this practice does not itself guarantee greater veracity of temperature assignments, since it rests on the assumption that temperature varies linearly with changes in the height of thermometric fluid. But this assumption cannot itself be verified, as that would require access to temperature in the world by some means independent of thermometry. Similar achievements, (seemingly) inextricably entangled with theory, may be seen at each further stage in the development of thermometric practice.

The moral of this case study is the historical contingency of thermometry, and thus of its results. At each stage in the development of thermometry, an advance in theory was required to extend measurement practice. Internal criteria of consistency and increased precision in the assignment of numerical values establish the new practice as an advance over the previous one. Yet, the application of these criteria is not empirically constrained. When one assumes that "temperature" (whatever it may be) varies linearly with changes in the height of the indicator column in an air thermometer, one is making an assumption both necessary for measurement progress and in principle non-empirical, since no independent access to "temperature," outside the behavior of the very devices and procedures under investigation, is possible: "*Prior to the construction of a thermometer, there is no thermometer to settle that question!*" (van Fraassen 2008, 126, emphasis in original). Chang (2004) argues that, in order to make sense of the "progress" exemplified by cases like these, we have to "look away from truth," and appeal only to historically contingent criteria for success (227)—"scientific progress … cannot mean closer approach to the truth" (228); "Truth, in the sense of correspondence to reality, is beyond our reach" (Chang 2007, 20). The delusion that one may evaluate the correspondence between our assignment of temperatures and the objective state of the world rests on the mistaken and "impossible god-like view in which nature and theory and measurement practice are all accessed independently of each other" (van Fraassen 2008, 139). Rather, the only relevant notion of "truth" for assessing the success of thermometry "rests first and foremost on coherence with the rest of the system" (Chang 2012, 242).

§2.2 Models and Calibration

Another kind of epistemic loop is found in synchronic measurement practice, where *models* play a constitutive role in determining measurement outcomes. The crucial concept here is *calibration*, the process of correcting a measurement device for inferred discrepancies between its readout and the target value. Calibration is a necessary feature of all sophisticated measurement, yet the process of calibration illustrates the ineliminable role of theoretical posits in the very assignment of quantity values in an act of measurement. When measuring, scientists do not (as one might naively suppose) read values directly from nature, rather they employ models of the interaction between measurement device and target system in order to "correct" the readout value to a final assigned value (Mari and Giordani 2014).

Tal (2014) illustrates this point through the example of the measurement of time, in particular coordinated universal time (UTC). The second is presently defined as 9,192,631,770 periods of the hyperfine transition between the two ground states of a caesium-133 atom at zero degrees Kelvin.[1] Models feature at every step of the process leading from devices that interact directly with caesium atoms to the UTC. First, it is impossible to probe caesium atoms at absolute zero, so the enumeration of hyperfine transitions output by a caesium clock must be corrected for this discrepancy. This, as well as other corrections, rely on models of the physical interaction between the device and the atom in order to infer the discrepancy between the actual state of the system and the idealized state referred to in the definition. Caesium clocks are too complex to run continuously, so their output is used to calibrate more mundane atomic clocks (301). Furthermore, the UTC itself is not identified with the output of any one clock; rather, it is calculated retrospectively by a weighted average over all participating atomic clocks, with weights determined by the degree of past fit between each clock and previous calculations of UTC (302–3).

The lessons of this example are analogous to those of epistemic iteration: measurement improvement appears to rest on internal standards of coherence rather than on correspondence with external quantities. The weighting procedure that leads to UTC, for instance, "promotes clocks that are stable relative to each other" (304). Success at achieving this stability indeed demonstrates "genuine empirical knowledge," but not knowledge in the first instance about a regularity in the objective world, but rather a regularity "in the behaviour of instruments" (327). Consequently, it is a "conceptual mistake" to think that "the stability of measurement standards can be analysed into distinct contributions by humans and nature" (328). On an extreme interpretation of this view, even computer simulation constitutes a form of measurement (Morrison 2009). The basic idea is that, once we grant the ineliminable role of models in measurement, it is a small conceptual step to accept that the aspect of measurement involving empirical contact with the world may be arbitrarily distant from that involving modeling (Parker 2017).

§3 Achieving Successful Measurement

For the remainder of this paper, I wish to grant the basic descriptive features of this account: both diachronically and synchronically, successful measurement involves epistemic loops. Nevertheless, I will argue, there is a form of measurement realism consistent with these loops; one on which the contingent, interest-relative, and theory-laden aspects of measurement may indeed be factored out, leaving the bare, objective facts about the world conveyed by successful measurement.

[1] Arguably, the process of establishing UTC is not measurement at all — since the length of the second is *defined* by caesium-133 transitions, it is not subject to empirical determination. The purpose of the project Tal examines is not to establish a value, as in paradigmatic cases of measurement, but rather to coordinate time-relevant activities across the globe with maximal precision. I set this concern aside for the discussion here, since Tal's analysis has been so influential in philosophy of measurement, and his conclusions concerning the model-mediation of measurement incontrovertibly reflect the practices of metrologists.

But what is "successful measurement"? For the purposes of discussion here, I take *measurement* to be any empirical procedure for assigning points (or regions) in a metric space to states of the world, where a *metric space* is any set of elements with a distance metric defined over it. This means, on the one hand, that I rule out degenerative forms of "measurement" that simply assign objects to categories, or place them in an order (the *nominal* and *ordinal* scales of Stevens 1946). On the other hand, I include measurement procedures that map states of the world into any geometrical space, not just the real line, so long as they have an assigned distance metric (siding with Suppes, et al. 1989, against Díez 1997); nevertheless, in the interests of simplicity, I will refer to these outcomes as "numerical" assignments, since they may be represented by vectors of real numbers. In line with Krantz et al. (1971), I take it that one can determine whether or not an empirical procedure constitutively requires the metric features of a geometrical space by analyzing whether these remain invariant across permissible transformations over the mapping into that space.[2]

I take *successful* measurement to exhibit two key features: *convergence* and *precision*. These features pose a significant challenge to the thoroughgoing coherentist.

§3.1 Convergence

Coherentists have emphasized the theory-ladenness of both diachronic and synchronic aspects of measurement refinement. However, a hallmark of sophisticated scientific measurement is its attempt to factor out the role of theory in measurement by employing different theoretical commitments to measure the same quantity. A measurement practice *converges* when procedures employing different theoretical commitments arrive at the same outcome.

For instance, in the early 20th century, a wide variety of phenomena were investigated, employing distinct methods and theoretical commitments, in the attempt to measure Avogadro's constant N_A, the number of particles in a mole of substance. Perhaps most well-known are Perrin's experiments on Brownian motion, which, in combination with Einstein's theoretical analysis, allowed an assignment of value to N_A. However, similar values were achieved by radically different means. For instance, Millikan was able to determine N_A by measuring charge of the electron through his oil drop experiments and dividing the Faraday constant (charge of a mole of electrons) by his result. Millikan's measurement relied on Stokes' theoretical analysis of the movement of spheres through a viscous fluid — insofar as Brownian motion was a factor, it was as a source of noise, not (as for Perrin) a source of evidence. Black body radiation and the blue

[2] For instance, consider two procedures for assigning real numbers to my students. On the first, I assign a number to each letter-type with which a student's name begins (e.g. A=1, B=3,…); on the second, I hold a meter stick up to each student and note their height. The former procedure is indifferent to the algebraic structure of the real line (letters do not add or subtract from each other systematically), and thus metric features of the real line are not invariant across alternative, equally permissible assignments of numbers (e.g. A=7, B=15,…). The second does make use of algebraic structure (as heights do "add" through concatenation), and thus metric features remain invariant across alternative assignments (Jamal is twice the height of Leslie, whether their heights are represented in inches or centimeters). So, on the present definition, the latter procedure is measurement, but the former is not.

of the sky are examples of other phenomena that, when combined with theoretical models of photon emission and diffraction respectively, allow alternate means of measuring N_A. Insofar as these procedures assign the same value to N_A, they converge.

I want to stress that the point being made here is *not* the traditional realist one, that these practices provide converging evidence for the particulate nature of matter, whether as "common cause" (Salmon 1984) or most likely hypothesis (Psillos 2011). Those arguments are instances of *abduction*, while I am interested in whether a stronger, non-abductive conclusion may be drawn from convergence. A better analogy is with the discussion of robustness in the modeling literature: a result is *robust* if it is obtained by a plurality of models that each make different simplifying assumptions (Weisberg 2006). The particulate nature of matter is not robust in this sense across different measurement practices, since it is assumed by all of them. However, the value of N_A is robust, since that value is not itself assumed, and is obtained with a great degree of agreement despite differences in the assumptions made by each measurement practice (and its supporting models). I claim that convergence toward this value provides robust, non-abductive evidence for an objective feature of the world.

This example is in no way exceptional: convergent measurement practices are rife across the sciences. Smith and Miyake, for instance, have investigated a number of examples. Thomson's convergent measurements of the charge of the electron employed a variety of different methods and assumptions (Smith 2001). Early attempts to measure the density of the interior of the earth likewise assumed a variety of different theoretical models (Miyake 2018). In more recent research, measurements of the constants that govern molecular vibration converge across spectroscopy, chemistry, thermodynamics, and femtochemistry (Smith and Miyake, *manuscript*). To pick an example from an entirely different area of science, measurements of the spectral sensitivity of mammalian retinal receptors employing psychophysical methods (extracting sensitivity curves from behavioral color matching experiments, as performed by Helmholtz in the late 19th century) converge closely with 20th century physiological methods (detecting rate of nerve firing in (e.g.) cow retinal tissue in response to single wavelength lights, Wandell 1995). In all of these cases, "What is being shown through the convergence of these measurements is that the discrepancies between the different measurements … are due to the particularities of the models being used" (Miyake, 2018, 336). In other words, convergence factors out model-sensitive features of measurement; in order for it to occur, "the empirical world has to cooperate" (Smith 2001, 26).

§3.2 Precision

Traditionally, measurement success was evaluated with respect to two features: accuracy and precision. *Accuracy* was degree of approach to true value, while *precision* was degree of specificity in the value provided. The considerations in §2 undermine the criterion of accuracy, since they show we have no independent access to "true values" and thus cannot use them as standards for evaluating measurement (Mari 2003). Nevertheless, we can still assess measurements for precision, since it may be defined operationally: a measurement is *precise* to the

extent that it returns the same result when performed repeatedly. The number of *significant figures* in a numerical assignment indicates the degree of measurement precision, since these characterize the size of the region within which repeated measurements fall.

Coherentists stress the fact that increased precision is a purely internal criterion for improving measurement. Here, however, I want to stress the way in which increased precision constitutes a qualitatively different, and more impressive, epistemic achievement than other forms of empirical success, such as qualitative prediction or improved coherence of classification. These qualitative achievements are subject to worries about semantic and theoretical holism: one may always succeed in classification, or correct qualitative prediction, by suitably redrawing the boundaries of one's theoretical concepts. As LaPorte (2004) argues, when faced with anomalies in the relationship between guinea pigs and prototypical rodents, or birds and dinosaurs, scientists face a *choice* whether to expand or contract their previous categories to include or exclude perceived outliers (a similar case is made by Slater 2017 for Pluto and planethood). Nothing about the prior conceptual framework itself forces this choice one way or another, nor do demands for internal consistency.

Measurement is different from mere categorization precisely because it maps states into a metric space. The crucial point to note here is that a metric space has *antecedent structure*: the distances between points on the real line, and the algebraic relationships between them, are fixed *before* we employ it to represent height or temperature or electric charge. This antecedent structure constrains the relationship between measurement outcomes, independently restricting our assessment of them as same or different, or converging or not, in a manner impervious to ad hoc revision. Increase in precision occurs when successive measurement practices are able to shrink distances (between repeated measurements within each practice) determined by the metric of the representing space. Thus, the metric of this space serves two functions: (i) it represents the distances between different measured quantities, but (ii) it also provides a directed metric for improving measurement of a single quantity, since it determines the distances between repeated measurements that characterizes their precision. Consequently, pace van Fraassen, attempts to increase precision are empirically constrained, since this directed metric for improvement can only be satisfied through the cooperation of nature: if nature is not sufficiently stable where we probe it, no choice, convention, or increased coherence can reduce the distances between our repeated attempts to measure it. Some examples will illustrate this point.

Consider, for instance, determinations of the boiling point of water. Chang (2004, Ch. 1) surveys the sequence of choice points in the early practice of thermometry leading to relative stability in the measurement of this temperature: what are the visual indicators of boiling, where should the thermometer be positioned, what should be the shape of the vessel holding the water, its material, etc.[3] Decisions on each of these points affect the relative stability in the thermometric reading, illustrating the naivety of a view on which

[3] The issue here is the phenomenon of "superheating," whereby water with relatively little dissolved gas, or in a flask with very small surface area, may be heated to a higher temperature without bubbling.

boiling point is a simple phenomena merely waiting to be observed. Nevertheless, in committing to represent the boiling point numerically, investigators subjected themselves to a criterion for success distinct from coherence. If the numbers assigned by thermometers within this-shaped vessels and that-shaped ones differ during phenomenologically similar bubblings, then the distance between those numbers provides a criterion of difference that must be respected if thermometric practice is to count as measurement. Restricting attention to those vessels that minimize distances between numerical outcomes is thus not a mere choice, or gerrymandering of the category "boiling," since it is forced upon the investigator by an antecedent metric for success.

Likewise, consider again the determination of UTC through the retrospective weighting of the comparison set of atomic clocks. For Tal, the success of this procedure is evidence for stability in our clocks, but not for any human-independent feature of the world. Nevertheless, UTC is constrained by the world in two distinct ways. First, through empirical contact with caesium atoms. While this contact is mediated by models, these models themselves are the result of convergent measurements of atomic phenomena through a wide variety of means, employing distinct theoretical assumptions. Second, the distance metric of the real line constrains the assessment of fit between clocks in the set. While the algorithm that weights them takes degree of internal agreement as the standard for higher weighting, the metrical structure of the space in which relative rates of the clocks are assessed ensures relative agreement cannot be stipulated, fudged, or gerrymandered. The clocks need to cooperate by performing stably enough that they may be compared with a high degree of precision, and this stable point remains tethered to a robust regularity in the world through checks with the convergent behavior of caesium.

While UTC is in some respects atypical (see footnote 1), these three features — internal coordination of outcomes, empirical checks, and directed improvement constrained by the real line — are features of scientific measurement in general. What Tal's discussion of the UTC obscures is the sheer number of empirical checks typically involved, and the strictness of the demands placed by conformity to the metric of improvement the measuring space provides. In official determinations of fundamental physical constants, convergence is demanded across *all* measurement procedures, as assessed by the law-governed interrelationship between physical quantities, and the degree of precision achieved illustrates the strictness of this demand. For instance, in late 19th century measurements of N_A by Perrin and e (charge of electron) by Thomson, only 2 to 3 significant figures were typically obtained within method, and convergence across methods often only agreed as to order of magnitude. By 1911, Millikan was measuring both e and N_A to 4 significant figures, and demonstrating that the models employed to calibrate the oil drop method converged closely with other aspects of physical theory (1911). As of 2014, N_A was being measured at upwards of 9 significant figures, and e upwards of 11 (Mohr et al. 2016).[4] In each case, the increase in precision has been constrained by the antecedent structure of the real line, and thus is not itself a matter of mere convention or coherence. Rather, the world must cooperate by remaining

[4] It is expected that after the 2018 26th General Conference on Weights and Measures, N_A and e will be fixed as constants to which other quantities may be referred during measurement.

sufficiently stable if such precision is to be possible; consequently, precise values constitute robust evidence for points of objective fixity in the world revealed through measurement.

§4 Conclusion: Fixed-Point Realism

Traditional scientific realism rests on an abductive inference from observed empirical success to presumed underlying causes. Successful measurement may certainly be used in such an inference, but I claim here that it non-abductively supports a more modest realism:

> **Fixed Point Realism** – values obtained through successful measurement veridically represent objective fixed points in the world, which may be exhaustively characterized by the pattern of distances that obtain between them in a metric space.

FPR is a form of *epistemic structural realism*. It differs from traditional realism insofar as it claims a veridical characterization of the world is possible independent of any particular theoretical description. Our theory of the nature of temperature or of state changes may change radically, yet the points of relative stability characterizing, e.g., boiling point of water, "absolute zero," freezing point of oxygen, etc., will stay robust across any such change, and that robustness may be represented by their relative positions within a numerical scale.

FPR differs from other flavors of structural realism in the type of structure to which it is committed. Structural realists typically focus on the rich mathematical structure of physical theory, and derivation or limit relations that hold between successive theories, e.g. Newton's laws are a limit case of relativistic mechanics (Worrall 1989). FPR commits itself only to *geometric* structure, i.e. the pattern of relative distances that obtain between points of stability as represented in a metric space. Just as our theoretical description of these stable points may change, so may our mathematical account of their relationship — if new mathematical physics fails to derive old equations as limit cases, this in no way jeopardizes the veridicality of this geometric structure.

Finally, FPR disagrees with coherentism, insofar as it asserts that the geometrical structure uncovered through acts of successive measurement obtains in the world independent of our practices. It does not deny the importance of epistemic loops for understanding the process of measurement. Nevertheless, it takes convergence in measured values to indicate that the points of stability they represent obtain independent of the theoretical commitments encapsulated in the models used for calibration. Likewise, it takes increased precision to constitute a criterion for measurement success over and above that of coherence, one that is only realized when the interest-independent world cooperates with us by remaining stable when we probe it.

Bibliography

Byerly, H., and V. Lazara (1973) "Realist Foundations of Measurement," *Philosophy of Science* 40:10–28.

Chang, H. (2004) *Inventing Temperature*, Oxford UP.

Chang, H. (2007) "Scientific Progress: Beyond Foundationalism and Coherentism," O'Hear (ed.) *Royal Institute of Philosophy Supplement* 61:1–20.

Chang, H. (2012) *Is Water H_2O?* Springer.

Díez, J. (1997) "A Hundred Years of Numbers: An Historical Introduction to Measurement Theory 1887–1990, part ii," *Studies in History and Philosophy of Science* 28:237–265.

Krantz, D., R. Luce, P. Suppes, and A. Tversky (1971) *Foundations of Measurement*, vol. 1, Dover.

LaPorte, J. (2004) *Natural Kinds and Conceptual Change*, Cambridge UP.

Mari, L. (2003) "Epistemology of Measurement," *Measurement* 34:17–30.

Mari, L., and A. Giordani (2014) "Modeling Measurement: Error and Uncertainty," in Boumans, Hon, and Petersen (eds.) *Error and Uncertainty in Scientific Practice*, Pickering & Chatto: 79–96.

Michell, J. (2005) "The Logic of Measurement: A Realist Overview," *Measurement* 38:285–294.

Millikan, R. (1911) "On the Elementary Electrical Charge and the Avogadro Constant," *Physical Review* 2:349–397.

Miyake, T. (2018) "Scientific Realism and the Earth Sciences," in Saatsi (ed.) *The Routledge Handbook of Scientific Realism*, Routledge: 333–344.

Mohr, P., D. Newell, and B. Taylor (2016) "CODATA Recommended Values of the Fundamental Physical Constants: 2014," *Review of Modern Physics* 88:035009.

Morrison, M. (2009) "Models, Measurement and Computer Simulation: The Changing Face of Experimentation," *Philosophical Studies* 143:33–57.

Parker, W. (2017) "Computer Simulation, Measurement, and Data Assimilation," *British Journal for Philosophy of Science* 68:273–304.

Psillos, S. (2011) "Moving Molecules above the Scientific Horizon: On Perrin's Case for Realism," *Journal for General Philosophy of Science* 42:339–363.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton UP.

Slater, M. (2017) "Plato and the Platypus: An Odd Ball and an Odd Duck – On Classificatory Norms," *Studies in History and Philosophy of Science* 61:1–10.

Smith, G. (2001) "J.J. Thomson and the Electron, 1897–1899," in Buchwald and Warwick (eds.) *Histories of the Electron*, MIT Press.

Smith, G., and T. Miyake (*manuscript*) "Realism, Physical Meaningfulness, and Molecular Spectroscopy"

Stevens, S. (1946) "On the Theory of Scales of Measurement," *Science* 103(2684):677–680.

Suppes, P., D. Krantz, R. Luce, and A. Tversky (1989) *Foundations of Measurement*, vol. 2, Dover.

Tal, E. (2013) "Old and New Problems in Philosophy of Measurement," *Philosophy Compass* 8/12:1159–1173.

Tal, E. (2014) "Making Time: A Study in the Epistemology of Measurement," *British Journal for Philosophy of Science* 67:297–335.

Wandell, B. (1995) *Foundations of Vision*, Sinauer.

Weisberg, M. (2006) "Robustness Analysis," *Philosophy of Science* 73:730–742.

Worrall, J. (1989) "Structural Realism: The Best of Both Worlds," *Dialectica* 43:99–124.

van Fraassen, B. (2008) *Scientific Representation*, Oxford UP.

Methodology at the Intersection between Intervention and Representation

Vadim Keyser[1]

Abstract: I show that in complex methodological contexts, representational and intervention-based roles require re-conceptualization. I analyze the relations between representation and intervention by focusing on the role of intervention in *mediating* representations. To do this, first I show how applied scientific practice challenges the simple distinction between representational and intervention-based roles of experiment/measurement. Then I discuss the complex interaction between representation and intervention applied to methodology in biomarker measurement.

1. Introduction

The relationship between intervention and representation is currently resurfacing in philosophy of science. Analytical treatments of the specific intersections between *representation* and *intervention* have recently been explored in Hacking (1983), Radder (2003), Heidelberger (2003), van Fraassen (2008), and Keyser (2017). These accounts analyze intervention-based experimental and measurement practice and the *consequences* for representing and model-building. Of particular interest in my discussion is that some of these accounts explicitly differentiate between representational and productive roles in scientific practice. For example, Heidelberger (2003) and van Fraassen (2008) discuss the representational and productive roles of instruments in experiment and measurement. In the former role, relations in a natural phenomenon are represented in an instrument (van

[1] California State University, Fresno. Email: vkeyser@csufresno.edu

Fraassen 2008, 94). In the latter role, instruments create new phenomena or mimetic phenomena, which resemble natural phenomena. Keyser (2017) takes the distinction between representation and production a step further to differentiate two types of experimental/measurement methodologies:

> When scientists measure/experiment they can *take* measurements, in which case the primary aim is to represent natural phenomena. Scientists can also *make* measurements, in which case the aim is to intervene in order to *produce* experimental objects and processes—characterized as 'effects'. (Keyser 2017, 2)

On Keyser's account 'taking a measurement' involves a scientist using a result in the context of theory to represent a given phenomenon (2017, 9-15). In contrast, 'making a measurement' involves setting up experimental conditions to produce a phenomenon—where that phenomenon can be realized in nature but it can also be a brand new phenomenon (Keyser 2017, 10). The difference between these two methodologies seems to be a matter of passive representation of a phenomenon vs. active intervention to produce a phenomenon. While the distinction between representation and intervention has been useful in classifying methodology in well-documented contexts like thermometry, microscopy, and cellular measurement, I argue that it falls apart in contexts where taking and making are *entangled*—such as in the context of biomarker measurement in the biomedical sciences.

In this discussion, I aim to show that in *complex methodological contexts*, representational and intervention-based roles require re-conceptualization. I analyze the *relations* between representation and intervention by focusing on the role of

intervention in *mediating* representations. In Section 2, I show how applied scientific practice challenges the simple distinction between representational and intervention-based roles of experiment/measurement. In Section 3, I discuss the complex interaction between representation and intervention applied to methodology in biomarker measurement.

2. Methodology at the Intersection between Intervention and Representation

In order to understand why the distinction between representation and intervention needs a multifaceted approach, it is important to be explicit about what it means to represent and intervene in scientific practice. In Section 2.1, I draw on van Fraassen (2008) to discuss representation and both van Fraassen (2008) and Keyser (2017) to discuss intervention. Then in Section 2.2, I show how applied scientific practice challenges the simplistic distinction between representational and intervention-based roles of experiment/measurement. I argue that the distinction between intervention and representation is less about *specific types of methodologies* in measurement/experiment and more about where one philosophically partitions the measurement *process*.

2.1. Representation and intervention

In experimental and measurement practice, representation has at least three important components: First, instruments or experimental contexts yield measurement values; Second, those values can only be interpreted within the context of a well-developed theory; and third, the relation between the measurement values and the phenomenon is determined by a user (e.g., experimenter). Van Fraassen (2008) provides

a rich characterization of representation in measurement and experiment, which requires careful analysis. Worth noting is that van Fraassen takes measurements to be a "special elements of the experimental procedure" (2008, 93-94). For my discussion the embeddedness of measurement in experiment is not important. I will focus on the roles or processes within measurement and experimental practice. But to do this, I will sometimes refer to 'measurement' and other times to 'experiment'. Van Fraassen's characterization focuses on interaction and representation in measurement:

> A measurement is a physical interaction, set up by agents, in a way that allows them to gather information. The outcome of a measurement provides a representation of the entity (object, event, process) measured, selectively, by displaying values of some physical parameters that—according to the theory governing this context—characterize that object. (2008, 179-180)

For van Fraassen, measurement interaction between an object of measurement and apparatus generates a physical outcome—the "measurement outcome" or "physical correlate of the measurement outcome"—, which provides information content about the target of measurement (2008, 143). The contents of measurement outcomes convey information about *what is measured* through the mediation of theory. Van Fraassen posits that theoretical characterization of measurement interaction requires 'coherence':

> The theoretical characterization of the measurement situations is required to be coherent with the claims about the existence of measurement outcomes, their relation to what is measured, and their function as sources of information. (2008, 145)

In short, the theory tells a coherence story about "how its outcomes provide information about what is being measured" (145). Furthermore, the information content is representational. Van Fraassen says, "The outcome provides a representation *of* the measured item, but also represents it *as* thus or so" (2008, 180). To understand how the representational relation works, it is important to refer to van Fraassen's 'representation criterion':

> The criterion for what sorts of interactions can be measurements will be, roughly speaking, that the outcome must represent the target in a certain fashion—, selectively resembling it at a certain level of abstraction, according to the theory—*it is a representation criterion*. (van Fraassen 2008, 141).

Two aspects of the representation criterion require explanation: First, the distinction between "target" and "outcome"; and second, the role of theory in the operation of measurement. I begin with the former. Van Fraassen makes a technical distinction between the target of measurement ('phenomena') and the outcome of measurement ('appearances'):

> Phenomena are observable, but their appearance, that is to say, *what they look like in given measurement or observation set-ups*, is to be distinguished from them as much as any person's appearance is to be distinguished from that person. (2008, 285)

For van Fraassen, phenomena are observable objects, events, and processes (2008, 283). He emphasizes that phenomena include all observable entities—whether observed or not (2008, 307). A given phenomenon can be measured in many different ways. The outcome of each measurement provides a perspective on a given phenomenon—meaning that the

content of measurement tells us what things *look like*, not what they *are like* (2008, 176, 182). The *content* of the measurement outcome is an appearance.

An important qualification is that for van Fraassen, a representation does not represent on its own. The scientist selects the aspects/respects and degrees to which a representation represents a target. This relation can be expressed as: Z uses X to represent Y as F, for purposes P.

Now that the target and outcome of measurement have been characterized, we can specify van Fraassen's role of theory in measurement. According to van Fraassen, "Measurement is an operation that locates an item (already classified as in the domain of a given theory) in a logical space, provided by the theory to represent a range of possible states or characteristics of such items (164). Three things are worth noting about van Fraassen's discussion of logical spaces. First, a logical space provides a multidimensional mathematical space that locates potential objects of measurement (2008, 164). By measuring we assign the item a location in a logical space. However, according to van Fraassen, it does not have to be on a real number continuum. As van Fraassen points out, items may be classified (by theory) on a range that is "an algebra", "lattice", or a "rudimentary poset" (2008, 172). Second, theoretical location depends on a "family of models" and not just an individual model (2008, 164). Third, an item is located in a "region" of logical space rather than at an exact point (2008, 165). Simply put, theory provides a classificatory system for what is measured. Importantly, theory is *necessary* for this type of classification. Van Fraassen says, "A claim of the form "This is an X-measurement of quantity M pertaining to S" makes sense *only* in a context where the

object measured is already classified as a system characterized by quantity M" (2008, 144 my emphasis).

We can summarize the above discussion into four conditions for van Fraassen's account of representation in measurement/experiment practice:

i. Physical Interaction Condition: The interaction between apparatus an object produces a physical correlate of the measurement outcome.

ii. Theoretical Characterization Condition: The content of the measurement outcome is given a location in a logical space, which is governed by a family of theoretical models. An item's location within a logical space can change in content and truth conditions as accepted theories change.

iii. Representational Content Condition: The content of a measurement outcome provides a selective representation of a given target of measurement (phenomenon). Because representations do not represent on their own, users and pragmatic considerations set the representational relation such that: Z uses X to represent Y as F, for purposes P.

iv. Perspectival Information Condition: Measurement generates appearances, which are public, intersubjective, contents of measurement outcomes. Appearances provide selective information about phenomena. Thus information from measurement tells us what something *looks* like and not what something *is* like.

Van Fraassen notes that measurement and experiment are not only limited to a representational role, they can take on at least two productive roles. First, instruments can produce phenomena that "imitate" natural phenomena. That is, carefully controlled conditions give rise to mimetic effects that are used by scientists in the context of theory to resemble natural phenomena (2008, 94-95). It is important to note that van Fraassen emphasizes that natural phenomena are phenomena that exist *independent of human intervention* (2008, 95). The second productive role of instruments is that they are used as "engines of creation" to produce or manufacture new phenomena. Van Fraassen is not explicit about whether or not the representational roles can smear with the productive roles. There is no reason to assume that these roles cannot be combined; but that requires explicit philosophical work to see *how*, which I develop in Section 3.

Keyser (2017) is explicit about the relationship between the representational and intervention-based roles in science. He discusses the *use* of intervention for developing causal representations. Scientists intervene, thereby manipulating causal conditions within a given measurement or experimental system, which he calls 'intervention systems', to produce some sort of "effect" (Keyser 2017, 9-10). According to Keyser, "Intervention systems consist of organized experimental conditions and as such the effects that emerge are often sensitive to changes in conditions" (Keyser 2017, 10). Once a given effect is produced it can be used in order to be informative about causal relations for theoretical model building.

Keyser (2017) also differentiates between the methodologies of taking measurements vs. making measurements. I interpret that taking measurements involves

three components: First, some instrument or experimental arrangement yields a qualitative or quantitative value; second, a 'theoretical representational framework'—which is just a body of models—is necessary in order to characterize that value according to parameters and relations between parameters; and third, a scientist sets up the resemblance relation between the measurement/experiment value and some aspect(s) of a phenomenon (Keyser 2017, 14-15). In contrast, when scientists make measurements they manipulate causal conditions—such as, preparatory, instrument, and background conditions—within an intervention system. This manipulation gives rise to some effect (Keyser 2017, 3-12).

There is something puzzling about Keyser's distinction between making vs. taking, if we apply the aforementioned conditions (i-iv): i. *Physical Interaction Condition*; ii. *Theoretical Characterization Condition*; iii. *Representational Content Condition*; and iv. *Perspectival Information Condition*. Namely, it seems that 'making measurements' is compatible with conditions i-iv, so it is not clear why there is a need for a distinction in methodological type, but rather just a difference in details for each condition. For example, when a measurement is made, there is a (i) *physical interaction* that occurs, but it is broader than just the instrument and object. The interaction can include "experimental conditions" (Keyser 2017, 3-5). The product of a made measurement is also amenable to (ii) *theoretical characterization*. Keyser emphasizes that theoretical characterization is necessary for experiment/measurement (Keyser 2017, 14); but he does not make the additional move to say that theoretical characterization is *part of the process* of making a measurement. That is, in order to make a measurement about an effect, one needs to also *characterize* that effect. Without the final

characterization, one is only dealing with the material conditions, which is an incomplete part of the measurement process. Keyser can accept that theoretical characterization is a necessary component of making a measurement. Otherwise, he risks offering a limited concept of 'making a measurement' that only applies to arranging the material components of the measurement process and nothing further.

The same challenge goes for (iii) *representational content* and (iv) *perspectival information*. An important component of the measurement process is to represent the relation between the produced effect and some aspect(s) of a phenomenon. For example, is this given effect a limited mimetic representation of a natural phenomenon or is it a brand new phenomenon? Without claims about what the effect is and its relation to objects, events, and processes in the world, 'making a measurement' is uninformative about part of the measurement process: the final value of the measurement outcome.

The aforementioned considerations question the need for a distinction between 'making' vs. 'taking'. One conclusion is that making uses the same components (i-iv), just with slightly different detail. But the other conclusion is a bit unsatisfying: making is really only about organizing the material components, which is an *initial* step in the measurement process, and it does not apply to later steps in measurement.

2.2. Dynamic relations between intervention and representation

I argue that the distinction between intervention vs. representation is less about *specific types of methodologies* in measurement/experiment and more about where to philosophically partition the *measurement process*. To make this point clear, I make two sub-points: 1) Measurement in the biological sciences offers complex and sometimes

blurred relations between instrument and object of measurement such that representation and production take on dynamic roles; 2) There is a difference between the act of measurement and the total process of measurement. I briefly describe (1) and (2).

On van Fraassen's (2008) and Keyser's (2017) characterizations of *representation* in measurement, the role of the instrument/apparatus seems to have an important mediating function. It may be the case that philosophical focus on case studies (e.g., thermometry, microscopy, cellular bio, and bacteria) that are instrument-intensive provide a certain support for an instrument-centric account of representation in measurement. Whether or not the necessary mediating role of instruments is an explicit part of both accounts, there is room to develop a richer philosophical view of the role of representation in the total measurement *process*. Without such philosophical development, we risk missing complex cases of measurement where intervention occurs side-by-side with representation. For example, in some cases of biological measurement, scientists use the organism to measure processes in that same organism but also to represent larger phenomena (Prasolova et al. 2006). For instance, mouse diets are manipulated in order to measure chromatin pattern changes. I characterize this as the mouse *constituting experimental conditions* that are being manipulated in order to measure some sort of process. The manipulation of conditions indicates an interventionist approach (or 'making' a measurement). Moreover, without manipulating the mouse's diet scientists would not be able to make a reliable measurement on chromatin structure at all. So the organism is not only being manipulated as part of the experimental/measurement set-up, it is a crucial part of that set-up. That is, without intervention, there is no reliable result. In addition to the organism being used as part of the measurement set-up, it also

serves as a physical *representation* of the dynamics of chromatin pattern change. That is, a given model organism can serve as a data model for a specific phenomenon of study—e.g., chromatin pattern in organism X. So, in this case the organism serves a dual function: it constitutes a set of experimental conditions to be manipulated and it serves as a physical representation of a phenomenon. Because of the dual function, this seems to be a case of both 'making' and 'taking' a measurement.

This brings me to sub-point (2). The total process of measurement is often complex in the biological sciences and requires multiple stages of intervening and representing. As mentioned in the model organism example representation and intervention are often *entangled*. Measurement is not merely putting an instrument up to something and waiting for a reading, which can be classified as an *act* of measurement. Measurement is also not merely creating effects out of material conditions. Measurement requires manipulation of conditions that is *used* in order to generate a representation. For example, identifying a mysterious fungus that is entangled with other fungus in a sample is an active process that requires both intervention and representation. One method is to take a sample and scrape it over a petri dish. What grows are spores that are passively deposited. But if common fungi were commingled with the mysterious fungi in the sample, and the common fungi grew faster, it would be impossible to identify the mysterious fungus. That is, coming back in a couple of weeks and seeing the petri dish covered with familiar species would lead to a false conclusion. Another way to perform the measurement (i.e. culture samples) is as follows. Take the samples and grind them up. Then sprinkle them into a petri dish. Put the dish under the microscope and, using a fine needle, pick out fragments of the mysterious fungus and transplant them to their own

dishes (Scott 2010). Once the fragments have been transplanted through this fine-grained intervention, each dish can be left to grow the colonies. The final dishes will offer visual representations that serve as data on the nature of the mysterious fungus. Notice here that intervention is a precursor to reliable representation.

Representation is not only reserved for the final instrument reading. It can also occur at other stages in the measurement process. Likewise, manipulation does not have to occur only at the earlier stages. For instance, organic matter can function as an instrument, like in the case of FourU thermometers, which are RNA molecules that act as thermometers in Salmonella (see Waldminghaus et al. 2007). Suppose that a scientist sets up an experiment to iteratively measure to what extent modifying RNA factors in FourU thermometers changes thermometer readings in Salmonella. In such a case the scientist could modify molecular factors and use the organic thermometers as temperature measures over many iterations, which would culminate in some sort of data model that organizes the relationship between molecular factors and FourU function. In such a case, there are multiple layers of intervention and representation.

The complex layering of intervention and representation is apparent in biomarker measurement in the biomedical sciences, where biological components serve as representations of disease conditions, but are also intervened on in order to make more reliable representations. I turn to this case study in the subsequent section.

3. Intervening in Representations and Representing Interventions

Biomarkers are used in biomedical measurement to reliably predict causal information about patient outcomes while minimizing the complexity of measurement,

resources, and invasiveness. A biomarker is an assayable metric—or simply, an indicator—that is used by scientists to draw conclusions about a biological process (De Gruttola et al. 2001). The greatest utility from biomarker measurement comes from their ability to help clinicians and researchers make conclusions with limited invasiveness. The reliance on biomarkers to make causal conclusions has prompted the use of 'surrogate markers'. These biomarkers are used to substitute for a clinically meaningful endpoint such as a disease condition. A major scientific methodological issue is that the use of multiple biomarkers will produce disagreeing results—and this is true even in the context of biomarkers that use similar biological pathways. To make methodological matters worse, theoretical representation is often not equipped to fill in the causal detail for each biomarker measurement. This amounts to an unfolding methodological puzzle about how to use intervention and representation in biomarkers to produce reliable measurements. My interest in this case study is not in solving the methodological puzzle, but rather in showing the *relations between intervention and representation* in such a complex case study. In this section, I discuss the complexity of intervention and representation in biomarker measurement to illustrate how intervention mediates the measurement process.

To understand the complex methodology in biomarker measurement it is important to detail the use and limitations of biomarkers. Some biomarkers are used as a substitute for some clinical endpoint. For instance, LDL cholesterol (LDL-C) is a biomarker that clinicians and physicians use to correspond to a clinical endpoint—e.g., heart attack. Moreover, the biomarker is associated with risk factors such as coronary artery stenosis, atherosclerosis, and angina pectoris. Katz (2004) argues that all biomarkers are candidates for 'surrogate markers', which can serve as substitutes for

clinical endpoints. That is, surrogate markers are reliable biomarkers that have a one-to-one correspondence with the disease condition such that they can be used to provide reliable predictive and causal information about a given clinical endpoint. There are a couple of points worth noting. First, notice that biomarkers and surrogate markers are being used as representations of a clinical endpoint. That is, to figure out the likelihood of developing a disease condition and to understand the risk factors associated with that disease condition, scientists use biomarkers that indicate information about the endpoint. This means that these physiological components can be used by clinicians and physicians to *represent disease conditions to respects and degrees*. The second point worth noting is that there are many biomarkers but limited surrogate markers and even more limited validated surrogate markers ('surrogate endpoints')—which are surrogate markers that are reliable in multiple contexts of interventions. The importance of this will be relevant shortly when I discuss the complexity of biomarker measurement. For our purposes, this means that most biomarkers in biomedical practice provide very limited representational information.

Surrogate markers are not passively used as physical representations of disease conditions. Their use is often more effective for representational purposes if there is a *mediating intervention*. For instance, surrogate markers can constitute "response variables". This is where a surrogate marker is manipulated in order to produce an effect that is relevantly similar to the effect with the same manipulation on the clinical endpoint. This means that an adequate surrogate must be "tightly correlated" with the true clinical endpoint; but it also means that any intervention on a surrogate marker must be tightly correlated with the intervention on the true clinical endpoint (Buyse et al. 2000). I

interpret this as a dual role for a reliable surrogate marker. It is to act as an epidemiological marker that *represents* some clinical endpoint but also to act as a responding variable that can be used in an *intervention* to causally influence the clinical endpoint. An example of the dual role of the surrogate marker is that high concentrations of LDL cholesterol (LDL-C) correspond to cardiovascular risk (Gofman and Lindgren 1950). But if a therapeutic interventions is used—such as, 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase inhibitors (statins)—that intervention can lower LDL levels, which in turn reduces cardiovascular disease (LaRosa et al. 2005).

So far I have presented the representational and intervention-based role of biomarkers. It is not straightforward to say that surrogate markers are *'made'* like an effect. But it is also not straightforward to say that surrogate markers constitute a *measurement outcome that is the final reading on an instrument*. These markers provide useful representational information *in the context* of an intervention. To add to the complexity of the relation between representation and intervention, biomarkers in the context of Alzheimer's measurement have added methodological steps. In Alzheimer's measurement there are different biomarkers, which are not correlated with each other and change with independent dynamics in the progression of Alzheimer's disease. So *each* of these biomarkers do not provide the same type of representation about the progression of Alzheimer's disease. Furthermore, scientists *only* understand the disagreement between each of these biomarkers in the presence of different interventions.[2] The different

[2] There has been much work recently on clinical biomarkers like: cerebrospinal fluid (CSF) tau, which is the primary component of neurofibrillary tangles; CSF 42-amino acid amyloid-β (CSF Aβ), which is the protein cleavage product believed to precipitate disease by forming neuron-damaging plaques; and amyloid plaques from PET scans.

interventions are in the form of drugs (e.g., bapineuzumab and solanezumab) and these interventions produce disagreeing representational results for the biomarkers. That is, the biomarkers respond differently to different interventions, which is methodologically problematic because it indicates that all of these biomarkers cannot be reliably tracking Alzheimer's progression in the same way. Interestingly, scientists systematically compare these disagreeing results to make reliable claims about Alzheimer's progression and treatment (Toyn 2015).[3] To simplify the method used, scientists track how interventions change properties of biomarkers and then they compare these amalgamated results with how interventions change behavioral/cognitive properties. This type of cross comparison allows scientists to eliminate biomarkers that do not track behavioral/cognitive improvement.

The structure of the methodological complexity in biomarker measurement can be partitioned as follows: 1) For a particular clinical endpoint, there are *limited physical representations* in the form biomarkers (or surrogate markers) which can be *used* to make representational and perspectival conclusions about the endpoint or risk factors associated with it; 2) *Scientists intervene in a process* from each of the biomarkers in order to track the relations between biomarkers and clinical endpoints; and 3) Such interventions

While the methodological story is beyond the scope of this discussion, there is a complex methodological point that is noteworthy for this discussion (Toyn 2015).

[3] To give a brief picture: The intervention of Bapineuzumab reduces levels of plaque assayed by Aβ PET and CSF tau, but not CSF Aβ; but Solanezumab *does not alter* levels of plaque assayed by Aβ PET and CSF tau but leads to a *reduction in* CSF Aβ. Cross comparison of the *intervention* mechanisms allows scientists to begin to make causal claims about which biomarkers are more reliable than others (Toyn 2015).

prompt *disagreeing results between the biomarkers*, which can 4) be amalgamated by researchers into further representations of the *relations between biomarkers and their clinical endpoints*. The above structural breakdown is merely *a* type of complex methodological process that can occur in biomedical measurement. It shows how interventions on physical representations (biomarkers) can produce other reliable representations. What is important to note about this analysis is the role of intervention in *mediating* further representations. In the case of biomarkers, intervention is necessary to test how close biomarkers are in their representations of clinical endpoints and also to other biomarkers. These representations not only represent the relation between the original biomarker and the clinical endpoint, but they also represent how a given intervention affects a given biomarker. As such, intervention paves the way for iterations of representations.

4. Concluding Remarks

In this discussion, I have analyzed the role of intervention in mediating representations by using examples from the biological and biomedical sciences. Characterizing intervention as a mediating factor in a larger methodological operation provides an important point about scientific practice. Representation and intervention are not neatly partitioned into contrasting methodologies. In fact, applied science often dictates the complex, and often smeared, philosophical concepts and methodologies. For this reason, I am proposing a *process* view of intervention and representation. This view opens up the diversity of relations between representation and intervention in a given experimental/measurement practice. While I have emphasized how intervention mediates

representation, there is more territory to explore about the mediating role of representation for intervention.

Work Cited

De Gruttola, V.G, Clax P, DeMets DL, et al. (2001). Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. Control Clin Trials 22:485–502.

Gofman, J.W., Jones, H.B., Lindgren, F.T., et al (1950). Blood lipids and human atherosclerosis. *Circulation* 2:161–178.

Hacking, I., (1983). *Representing and Intervening*, Cambridge: Cambridge University Press.

Heidelberger, M. (2003). Theory-ladenness and scientific instruments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 138–151). Pittsburgh, PA: University of Pittsburgh Press.

Katz, R. (2004). Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* 1:189–195. doi: 10.1602/neurorx.1.2.189

Keyser, V. (2017). Experimental Effects and Causal Representations. *Synthese*, SI: Modeling and Representation, pp. 1-32.

LaRosa, J.C., Grundy, S.M., Waters, D.D., et al. (2005). Intensive Lipid Lowering with Atorvastatin in Patients with Stable Coronary Disease. *New England Journal of Medicine* 352:1425–1435. doi: 10.1056/NEJMoa050461

Prasolova L.A., L.N. Trut, I.N. Os'kina, R.G. Gulevich, I.Z. Pliusnina, E.B. Vsevolodov,

I.F. Latypov. (2006). The effect of methyl-containing supplements during pregnancy on the phenotypic modification of offspring hair color in rats. *Genetika*, 42*(1)*, 78-83.

Radder, H. (2003). Technology and theory in experimental science. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 174–197). Pittsburgh, PA: University of Pittsburgh Press.

Toyn, J. (2015). What lessons can be learned from failed Alzheimer's disease trials? *Expert Rev Clin Pharmacol* 8:267–269. doi: 10.1586/17512433.2015.1034690

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.

Waldminghaus, T., Nadja H., Sabine B., and Franz N. (2007). FourU: A Novel Type of RNA Thermometer in Salmonella. *Molecular Microbiology* 65 (2): 413–24. https://doi.org/10.1111/j.1365-2958.2007.05794.x.

Philosophy of Science (forthcoming)
v1.2 (as of 9/15/18)
Please cite published version

Are Emotions Psychological Constructions?

Charlie Kurth
Department of Philosophy
Western Michigan University

Abstract: According to psychological constructivism, emotions result from projecting folk emotion concepts onto felt affective episodes (e.g., Barrett 2017, LeDoux 2015, Russell 2004). Moreover, while constructivists acknowledge there's a biological dimension to emotion, they deny that emotions are (or involve) affect programs. So they also deny that emotions are natural kinds. However, the essential role constructivism gives to felt experience and folk concepts leads to an account that's extensionally inadequate and functionally inaccurate. Moreover, biologically-oriented proposals that reject these commitments are not similarly encumbered. Recognizing this has two implications: biological mechanisms are more central to emotion than constructivism allows, and the conclusion that emotions aren't natural kinds is premature.

This paper challenges the psychological constructivist account of emotions that is gaining prominence among neuroscientists and psychologists (e.g., Barrett 2017, 2012, 2009; LeDoux 2015; Russell 2004). According to constructivism, emotions result from projecting culturally-fashioned concepts onto felt affective episodes. Fear, for instance, just is a feeling of negative arousal as viewed through the lens of one's folk concept FEAR. This proposal is novel in taking felt experience and cognitive projection to be essential elements of what emotions are. Moreover, while constructivists acknowledge that there's a biological dimension to emotions (e.g., neural mechanisms are responsible for generating the conscious feelings that we project our emotion concepts on to), they deny that emotions are, or necessarily involve, anything like an affect program. Thus, constructivism is philosophically significant in two ways. First, in denying an essential role for biological mechanisms, it challenges influential, affect-program-oriented accounts of emotion (e.g., Scarantino & Griffiths 2011; Ekman & Cordaro 2011). Second, in understanding emotions as projections of folk emotion concepts, it takes emotions to be social-psychological constructions, not natural kinds.

But despite constructivism's appeal among cognitive scientists, the role that it gives to felt experience and folk concepts leads to an account of emotion that's both extensionally inadequate and functionally inaccurate. Moreover, biologically-oriented proposals that reject constructivism's problematic commitments are not similarly encumbered. Recognizing all this reveals that an adequate account needs to give greater place to the biological mechanisms that underlie emotions than constructivism allows. This, in turn, suggests that the constructivists' conclusion that emotions are not natural kinds is premature.

1. Psychological Constructivism and Its Appeal

Constructivism sees emotions as having two elements: a felt affective experience and a cognitive projection or labeling. Taking these in turn, the felt experience component—or "core affect" as it's often called—is a neurophysiological state that manifests as a consciously experienced combination of valence (i.e., feeling good or bad) and arousal (i.e., feeling activated or deactivated) (Barret 2006: 48; Russell 2004; LeDoux 2015: 226-232). Importantly, constructivism's focus on core affect looks just to the amalgamated *experience* of these two components—valence and arousal. What *causes* this felt experience is irrelevant to the nature and individuation of emotions. In fact, and as we will see, allowing that particular sensations (instances of core affect) can be produced by a range of distinct neural circuits or somatic events is taken to be a point in favor of the constructivist proposal.

Given this account of the felt dimension, constructivism maintains that "discrete emotions emerge from a conceptual analysis of core affect. Specifically, the experience of feeling an emotion…occurs when conceptual knowledge about emotion is brought to bear to categorize a momentary state of core affect. … [These] [c]ategorization processes enact the rules, [that guide] the emergence of an emotional episode" (Barrett 2006: 49; also LeDoux 2015: 225-232). This talk of "conceptual analysis," "conceptual knowledge," and "categorization" should be understood thinly.

The underlying process needn't involve some full-fledged, conscious judgment. Rather, all that's necessary is an unconscious or implicit recognition that one's sense of one's situation, and one's felt physiological state, fall under a particular folk emotion concept.

These emotions concepts, in turn, should be understood as folk theories or culturally-shaped behavioral scripts that detail the nature and function of the particular mental states picked out by specific emotion labels ('fear,' 'joy,' 'anger,' etc.). Moreover, the fact that folk emotion concepts engage these folk theories and behavioral scripts entails that the projecting of a particular label onto an instance of core affect not only imbues one's situation with the associated, emotionally-colored meaning, but also shapes one's subsequent thoughts, physiological responses, and behaviors (Barrett 2012; LeDoux 2015).

Formalizing this a bit, we can see psychological constructivism as committed to four theses:

> (PC1) Each emotion type/category is constituted by the projecting of a specific folk emotion concept (e.g., FEAR, JOY) onto a felt affective experience.
>
> (PC2) Token emotion episodes (e.g., a given instance of fear) are cognitive acts where one (implicitly) labels an occurrent conscious feeling with a particular folk emotion concept and so comes to see the feeling through the lens of that concept.
>
> (PC3) There is no unique (set of) neural circuit(s) or psychological mechanism(s) responsible for the conscious feelings that get categorized with particular folk emotion concepts.
>
> (PC4) The act of labeling a feeling with a particular folk emotion concept affects one's subsequent thoughts, physiological responses, and behaviors.

According to its advocates, much of constructivism's appeal lies in its explanatory power. In comparison to more biologically-oriented theories, it provides a better explanation of empirical research on the biological mechanisms and correlates associated with emotions (e.g., neural circuits, patterns of physiological change, and expressive behavior). Since the discussion that follows will build

from the contrast between constructivism and competing biologically-oriented theories (BTs), it will be useful to briefly sketch the BT approach and the constructivists' case against it.

As a generalization, BTs maintain that emotions are, or necessarily engage, affect programs—that is, largely encapsulated systems that automatically prompt stereotyped patterns of physiological changes, expressive behavior, motor routines, attentional shifts, and forms of higher-cognitive processing in response to (evolutionarily-relevant) threats and opportunities. So, for example, fear is (or essentially involves) an affect state that consists of automatically engaged tendencies for inter alia increases in arousal, narrowing of attention, and the cueing of fight/flight/freeze behavior in response to the perception of some danger.

But since BTs take affect programs to be essential (even identical) to emotions, constructivists argue they cannot explain two well-documented sets of findings.[1]

> (F1) One can feel a given emotion without engaging what science suggests is the best candidate for its underlying biological drivers (or their correlates)—e.g., activation of particular neural circuits, a distinctive physiological response, characteristic expressive behavior.
>
> (F2) The relevant biological drivers/correlates can be engaged though one does not report feeling the associated emotion.

So, for instance, though the central nucleus of the amygdala (CeA) is thought to be central to fear, research shows both that individuals will report being afraid when the CeA is not engaged (F1), and that the CeA can be active though individuals report not feeling fear (F2).

BT proponents have sought to address these explanatory limitations by insisting that we must narrow our understanding of what, say, FEAR is. More specifically, they maintain that the folk emotion concepts that the above research relies on (in, e.g., the self-reports of emotions (not) felt) are too

[1] See, e.g., Barrett 2012 for a review of the relevant empirical work.

coarsely grained for scientific investigations like these. The BT advocates' expectation is that a more refined account of what 'fear' refers to will reduce, even eliminate, dissociations of the sort noted above (e.g., Scarantino & Griffiths 2011; Kurth 2018). But constructivists respond that any effort to narrow or otherwise refine our emotion concepts along these lines will result in an account of (e.g.) fear that is troublingly stipulative or excessively revisionary with regard to our ordinary understanding of these emotions (Barrett 2012: 415-6; LeDoux 2015: 234).

Two aspects of this debates are particularly important for our purposes. First, central to the constructivist complaint is the move to take a failure to accommodate our *ordinary emotion talk* as the standard for what counts as stipulative or excessively revisionary account. Second, given our ordinary emotion talk as the standard, the above four theses appear to give constructivism the resources and flexibility it needs to explain not just (F1)-(F2), but also the richness and cultural variation of emotional life more generally (e.g., Barrett 2012, 2009). However, I will argue that investigating the extensional adequacy and functional accuracy of constructivism's core theses provides us with reason to doubt each of (PC1)-(PC4).

2. Is Constructivism Extensionally Adequate?

As we've seen, a central feature of the debate between constructivism and BTs is the charge that BTs cannot accommodate dissociation data without committing to a stipulative or excessively revisionary account of what emotions are. In what follows, I give three examples that suggest constructivism faces a similar problem. More specifically, a closer look at the constructivists' dual claim that emotions are *cognitive labelings* of *felt experiences* reveals that the account is both under- and over-inclusive with regard

to our ordinary understanding of things like: what emotions are, when we experience them, and how they differ from moods, feelings, and other categories of affect.[2]

First consider the constructivist's commitment to understanding emotions as felt experiences—that is, changes in core affect that we're consciously aware of. An implication of taking felt affective experience as essential to being an emotion is that it rules out the possibility of unconscious emotions. Some constructivists appear to embrace this result. For instance, Joseph LeDoux maintains that claims about unconscious emotions are "oxymoronic" (2015: 234; also, 19). But LeDoux's acceptance of this implication aside, the thought that there cannot be unconscious emotions fits poorly with our everyday experiences and our ordinary emotion talk.

For instance, if there aren't unconscious emotions, then how do we explain situations where we don't realize that we were (say) afraid until *after* the danger has passed? Pressing further, notice that we not only regularly speak of unconscious emotions, but also appeal to them in order to explain our behavior. For example, we say things like, "Bill won't discuss the book he is working on. He says it's not ready yet—but he doesn't realize that he's really just afraid about getting negative feedback." While ordinary talk like this is easy to make sense of on the assumption that Bill is unconsciously fearful, such an explanation isn't available to a constructivist like LeDoux—our ordinary talk to the contrary, Bill isn't unconsciously afraid, but rather experiencing some other psychological blockage.

But the constructivists' trouble with unconscious emotions runs deeper—the case for their existence also has empirical support. For instance, recent experimental work has shown that subliminally presented emotion faces can produce affective responses that bring emotion-specific behaviors *even though* the subject denies feeling an emotion. In particular, subliminally presented happy

[2] Thus the strategy I employ here—one that *grants* constructivists' their criterion for assessing when an account is excessively revisionary—is distinct from standard defenses of BTs noted in §1.

faces bring increased "liking" behavior (e.g., greater consumption of a novel beverage), while subliminally presented angry faces have the opposite result (Winkielman et al. 2003; also, Kihlstrom 1999). Since these patterns of behavior mesh with our understanding of both joy as an emotion that tends to increase interest/engagement, and anger as an emotion that brings avoidance/rejection tendencies, these results are taken as evidence of unconscious emotions.

While the constructivist might try to pass these findings off as cases where unconscious changes in core affect (not emotion) produce the behaviors, the plausibility of the proposal is undercut by the fit we find between the subliminally presented happy (angry) face, the resulting liking (avoidance) behavior, and *our ordinary understanding* what happiness (anger) involves (Winkielman et al. 2005). The upshot, then, is that constructivism's insistence that felt changes in core affect are *essential* to what emotions are has revisionary implications with regard to our ordinary (and scientific) understanding of emotional life.

But even if we're willing to grant that our talk of unconscious emotions is merely metaphorical—an elliptical way of talking about some non-emotion form of (unconscious) affect—the constructivist's second core commitment brings additional problems. In particular, the claim that emotions are the product of our cognitive labelings/projections makes facts about when we are experiencing an emotion—and what emotion it is—too sensitive to random situational features and framing effects. To draw this out, consider the following case.

> Coffee. I order a cup of decaf coffee and sit down to read a magazine cover story about Trump's latest foreign policy provocations. But unbeknownst to me, the barista confuses my order and I get a cup of regular coffee. As the caffeine works its way into my system, it brings a (consciously experienced) change in my arousal. As a result, I start reading the article with jittery attentiveness.

Given the scenario, it seems my jittery, attentive reading is best understood as a bout of caffeine-induced hyperactivity. But notice: there's nothing in the constructivist account to rule out the

possibility that I'm actually having an emotional experience—I'm afraid. After all, on the constructivist account, this experience could be a change in core affect that I've (implicitly) labeled 'fear.' While that possibility alone seems odd (to my ear, at least, the case is best understood as emotionless hyperactivity, not fear), there's more trouble.

To draw this out, consider the constructivist's likely response to the case. Given the setup, she would likely maintain that whether this is an instance of fear depends on whether I see it that way—what sort of meaning do I attribute to my situation (e.g., Barrett 2017: 126; 2012: 419-420; 2009: 1293)? For instance, if I assent to the barista's remark that I seem really uneasy about the article that I'm reading, then—by (implicitly) labeling my behavior through my assent—I imbue my situation with the meaning carried by my FEAR concept. I am, therefore, feeling fear. While this move might seem to allow the constructivist a way to account for the case, it comes at a high cost. For notice, had the barista instead said something like, "Whoops, I messed up and gave you regular, not decaf—no wonder you're so hyper," I'd likely assent to that too. And so I wouldn't be afraid—just hyperactively aroused.

But that's odd. Our ordinary thinking about emotions suggests that whether I'm experiencing a particular emotion, and what emotion I'm experiencing, should *not* be so sensitive to random situational features like what questions the barista—or anyone for that matter—just happen to ask me. To be clear, the claim here is not that emotions are immune to situational and contextual factors. Rather, the point is that on the constructivists' account emotions turn out to be *too* sensitive to them. The radical situational sensitivity entailed by constructivism makes it not only too easy to experience an emotion, but also ties facts about what emotion we're experiencing to irrelevant situational factors.

Together, the difficulties raised by unconscious emotions and incidental situational features call the extensional adequacy of the constructivist account into question and do so in a way that

pinpoints the commitments of (PC1) and (PC2) as the source of the trouble—after all, these claims posit feelings of core affect and projections of folk concepts as essential to what emotions are. Of equal note is the fact that biological theories are less vulnerable to these difficulties. For one, irrelevant situational features should have less influence on what emotion one happens to experience since, according to BTs, emotions are (or are principally driven by) affect programs, not contextualized cognitive labelings. Moreover, since affect programs are things that can operate below that level of conscious awareness (Kurth 2018), taking emotions to be driven by affect programs provides BTs with the resources needed to explain unconscious emotions.

While the above discussion raises worries about the first two constructivist theses (PC1-PC2), it also provides the makings for worries about the third. In particular, because constructivism denies (via PC3) that emotions are underwritten by affect programs, it has trouble making plausible distinctions between emotions and similar states like moods. To draw this out, notice that the coffee case from above can be easily extended to show that constructivism makes it too easy to flip between moods and emotions. All we need to do is substitute "being in a worried mood" for "hyperactive" in the presentation of the case. Once we do this, we see that mere changes in the question the barista asks me can change whether I'm worried (a mood) or afraid (an emotion).

So we again see that constructivism has problematic explanatory limitations—this time with regard to preserving the thought that there's a substantive difference between moods and emotions. On the constructivist account, this distinction is just a matter of how we happen to label our felt experiences. While some constructivists appear willing to accept this conclusion (e.g., Barrett 2017, 2009), it highlights another place where the constructivist proposal has revisionary implications—after all, moods and emotions are generally thought to be *distinct* forms of affect (e.g., Ben-Ze'ev 2000: Chap. 4). Moreover, here too we have a difficulty that's easily avoided by biological accounts. Since

BTs take emotions to be (driven by) affect programs, they can appeal to the engagement of these mechanisms as the basis for the emotion/mood distinction (e.g., Kurth 2018; Wong 2017).

Stepping back, then, although constructivism purports to be less stipulative with regard to capturing our ordinary understanding of emotions, the above examples call this into question. For starters, the constructivists' commitment to (PC1)-(PC3) has revisionary implications for our ordinary understanding of what emotions are, when we experience them, and how they differ from moods. Moreover, we have also seen that biologically-oriented accounts—in eschewing this trio of problematic theses—are better equipped to provide a plausible account of these features of our everyday emotion talk.

3. Is Constructivism Functionally Accurate?

The challenges to the constructivist picture extend beyond concerns about its extensional adequacy. The account also makes predictions about how projecting emotion concepts onto felt experience should shape subsequent behavior that are poorly supported by the empirical record. Two examples will draw this out.

First consider emotion misattribution research. In this work, a feeling that is typically associated with a particular emotion (e.g., feelings of unease and anxiety) is subtly induced, but the individual is lead to believe they are not, in fact, experiencing that emotion but rather something else (e.g., the effects of caffeine). Constructivism predicts (via PC4) that individuals in these experiments should display different behaviors depending on whether they are in the control or misattribution conditions. For instance, individuals led to believe that the unease they're feeling is not anxiety, but something else (caffeine) should display diminished anxiety-related behaviors in comparison to controls who were not misled about their unease. But on this score, the experimental findings are decidedly mixed.

First, while there is a sizable body of findings showing misattribution manipulations attenuate subsequent emotion-related behavior, there is also a sufficiently large set of non-confirmations to raise concerns. For instance, while some research on public speaking anxiety suggests that attributing unease to a pill you just took rather than anxiety about a public talk you must give leads to a reduction in anxiety-related behaviors—stuttering, apprehension, and the like (Olson 1988), other studies have failed to find any differences in these behaviors (Slivkin & Buss 1984; Singerman, Borkovec & Baron 1976).

Moreover, even in cases where emotion-related behavior is reduced in the manipulation condition, it's not clear how much support this brings to the constructivist. This is because it's often unclear whether the reductions in emotion-specific behavior are (i) the result of the misattribution or (ii) a consequence of directing subjects' attention away from the emotion eliciting stimuli (for a review, see, e.g., Reisenzein 1983). This potential confound is problematic for constructivists since only possibility (i) provides direct support for the claim of (PC4)—namely, that the act of labeling *itself* affects subsequent behavior.

The second problematic set of results comes from work in political science. This research investigates how negative emotions shape public policy decision making among voters (e.g., MacKuen et al. 2010; Brader et al. 2008; Valentino et al. 2008). The core hypothesis of this research is that negative emotions (especially, anger and anxiety) affect subsequent behavior in different ways. In particular, anger—as a response to challenges to what one values—should tend to bring behavior geared toward defending the threatened values. By contrast, since anxiety is a response to uncertainty, it should tend to bring caution and information gathering aimed helping one work through the uncertainty one faces.

To test these predictions, the experimental set up works as follows. First, individuals are asked to read a (fake) news story designed to provoke anger or anxiety by challenging the individuals' pre-existing views about contentious policy issues like immigration, affirmative action, and economic policy. After reading the story, the participants are given the opportunity to use a website containing links to additional information, both for and against, the policy issue at hand. They are also asked how the original news story they read made them feel (e.g., angry, anxious). So by tracking what kinds of information the participants looked at through the website, experimenters can identify differences in how the anger and anxiety provoked by the story shaped subsequent behavior.

In the present context, these experiments allow us to test a pair of predictions that follow from the constructivist theses (PC1) and (PC4):

> (P1) Labeling felt experiences with distinct folk emotion concepts should bring different patterns of behavior.
>
> (P2) The behaviors that result from labeling a felt experience with a particular concept should map to our folk understanding of the emotion in question.[3]

More specifically, given (P1) and (P2), we should see different behaviors based on whether the participants in the experiment label their emotion 'anger' or 'anxiety' (P1). Moreover, the different behaviors should map to the above, ordinary understanding of these emotions—e.g., angry individuals should look for information that helps them defend their preferred policy position, while anxious individuals should engage less in motivated inquiry and more in open-minded forms of investigation (P2).

[3] As evidence of constructivism's commitment to these predictions, consider Lisa Feldman Barrett's comment that "when a person is feeling angry…she has categorized sensations from the body and the world using conceptual knowledge of the category 'anger'. As a result, that person will experience an unpleasant, high arousal state as evidence that someone is offensive. In fear…she will experience the same state as evidence that the world is threatening. And, *either way, the person will behave accordingly*" (2009: 1293, emphasis added).

However, whether we find support for these predictions turns—surprisingly—on what the policy issue used in the experiment was. More specifically, in experiments where the policy question that was challenged by the fake news story concerned immigration, the results fit poorly with constructivism's predictions. That is, participants behaved in the same angry way regardless of whether they reported feeling anger or anxiety (Brader et al. 2008). By contrast, if the policy issue at hand concerned affirmative action or economic policy, the results are more in line with (P1)-(P2): anger and anxiety provoked by the news stories not only brought different patterns of behavior, but the resulting behaviors mesh with our ordinary conception of how these emotions function (MacKuen et al. 2010; Valentino 2008).

While this second set of results might appear to be good news for constructivists, the trouble lies in explaining why we get the different results between the immigration and affirmative action/economic policy experiments. After all, other than the content of the issue at hand, the experimental designs were *identical*. In response, the constructivist might argue that content and context matter (e.g., Barrett 2012, 2009): the similar behaviors that subjects display in the immigration version of the study suggest that the cultural scripts associated with 'anger' and 'anxiety' are highly sensitive to negative stereotypes about minorities. More specifically, the thought would be that there's something about the combination of immigration debates and racial stereotypes that changes the standard behavioral scripts associated with 'anger' and 'anxiety' so that, while they *typically* generate different behaviors, they *now* bring the same ones.

But setting aside concerns about the ad hoc nature of this proposal, without more of a backstory, it's unconvincing. After all, affirmative action debates are *also* framed in racial stereotype provoking ways. So here too we should see anger and anxiety generating similar patterns of behavior. But we don't.

Moreover, notice that, on this front, biological accounts have an easier time explaining the experimental findings. For instance, as one possibility, the BT advocate could argue that only participants in the immigration study are likely to be experiencing *both* anger and anxiety: anger about the harms immigrants will bring and anxiety given their uncertainty about the likelihood of these harms. Given this, the BT advocate could then add two claims about what happens when both these emotions are engaged. First, since anger is a more powerful emotion than anxiety, it tends to win out with regard to shaping individuals' subsequent behavior. Second, given the high degree of overlap in the felt experiences produced by the anger and anxiety affect programs (e.g., both bring increased, negatively valenced arousal), when prompted to state what emotion they are feeling, some subjects happen to interpret their feelings as anger, while others see it as anxiety. Thus, the BT advocate can explain both why we get mixed results when subjects are prompted to state what emotion they are feeling and why, despite these differences in self-reports, the individuals nonetheless respond with behavior characteristic of anger, not anxiety. Moreover, because this proposal allows anger to drive behavior *regardless* of how subjects happen to label it, the explanation is unavailable to constructivists.

All told, we have two independent sets of experimental findings showing (at best) equivocal support for constructivism's predictions about how projecting emotion concepts onto felt experience should shape subsequent behavior. Moreover, we've also learned that more biologically-oriented accounts are better able to handle the experimental findings we've reviewed.

4. Conclusion: Emotions, Biology, and Natural Kinds

As we've seen, constructivism's purported advantage over more biologically-oriented theories lies is its ability to better explain the richness and diversity of emotional life (§1). But we have also seen that a crucial premise in this argument is the move to take accommodating our ordinary emotion talk as the standard for assessing a theory's explanatory power. Not only are there familiar problems for

adopting such a standard (e.g., Scarantino & Griffiths 2011, Kurth 2018), but—even if we accept it—we've learned that there's trouble for constructivism. In particular, the explanatory "success" constructivism secures come by way of a highly revisionary account of what emotions are, when we experience them, how they differ from moods, and the way that they shape behavior (§§2-3). Moreover, our critical observations also implicate the four constructivist theses (PC1-PC4) as the source of these difficulties. Thus it's not surprising that more biologically oriented proposals—accounts that reject these commitments—do not face similar explanatory limitations.

Taken together, then, the arguments of this paper suggest a pair of larger lessons. First, even if we agree that constructivists are correct about what the relevant standard for assessing a theory of emotion is, we've learned that an adequate account must give greater place to the biological mechanisms that underlie emotions than constructivism allows. This, in turn, indicates that the constructivists' conclusion that emotions are not natural kinds is premature. After all, if we must posit something like an affect program in order to (i) explain everyday talk and empirical findings about unconscious emotions, (ii) capture the thought that emotional experience is not radically sensitive to random situational features, and (iii) accommodate research regarding how emotions shape behavior, then we have evidence that (at least some) emotions are underwritten by mechanisms that make them plausible candidates for being natural kinds.

References

Barrett, L. 2017. *How Emotions Are Made*. New York: Houghton Mifflin Harcourt.

---. 2012. "Emotions Are Real." *Emotion* 12: 413-429.

---. 2009. "Variety is the Spice of Life." *Emotion and Cognition* 23: 1284-1306.

---. 2006. "Emotions as Natural Kinds?" *Perspectives on Psychological Science* 1: 28-58.

Ben-Ze'ev, A. 2000. *The Subtlety of Emotions*. Cambridge.

Brader, T. et al. 2008. "What Triggers Public Opposition to Immigration?" *American Journal of Political Science* 52: 959-978.

Ekman, P. & D. Cordaro. 2011. "What is Meant by Calling Emotions Basic." *Emotion Review* 3: 364–370

Kihlstrom, J.F. 1999. "The Psychological Unconscious." In L.A. Pervin & O.P. John (Eds.), *Handbook of Personality* (2nd ed., pp.424–442). New York: Guilford Press.

Kurth, C. 2018. *The Anxious Mind.* MIT Press.

LeDoux, J. 2015. *Anxious.* New York: Viking.

MacKuen, M. et al. 2010. "Civil Engagements," *American Journal of Political Science* 54: 440–458.

Olson, J. 1988. "Missattribution, Preparatory Information, and Speech Anxiety" *Journal of Personality and Social Psychology* 54: 758-767.

Reisenzein, R. 1983. "The Schachter Theory of Emotion" *Psychological Bulletin* 94: 239-264.

Russell, P. 2004. "Core Affect and the Psychological Construction of Emotion. *Psychological Review* 110: 145–172

Scarantino, A & P. Griffiths. 2011. "Don't Give Up on Basic Emotions" Emotion Review 3: 1-11.

Singerman, K. et al. 1976. "Failure of a 'Misattribution Therapy' Manipulation with a Clinically Relevant Target Behavior" *Behavior Therapy* 7: 306-316.

Slivken, K. & A. H. Buss. 1984. "Misattribution and Speech Anxiety" *Journal of Personality and Social Psychology* 47: 396-402.

Valentino, N. et al. 2008. "Is a Worried Citizen a Good Citizen?" *Political Psychology* 29: 247–73.

Winkielman, P. et al. 2005. "Unconscious Affective Reactions to Masked Happy versus Angry Faces Influence Consumption Behavior and Judgments of Value." *Personality and Social Psychology Bulletin* 121-135.

Wong, M. 2017. "The Mood-Emotion Loop" *Philosophical Studies* 173: 3061-3080.

Symposium: ***Bridging the Gap Between Scientists and the Public*****, PSA 2018**

How trustworthy and authoritative is scientific input into public policy deliberations?[i]

Hugh Lacey
Swarthmore College / University of São Paulo

Abstract: Appraising public policies about using technoscientific innovations requires attending to the values reflected in the interests expected to be served by them. It also requires addressing questions about the efficacy of using the innovations, and about whether or not using them may occasion harmful effects (risks); moreover, judgments about these matters should be soundly backed by empirical evidence. Clearly, then, scientists have an important role to play in formulating and appraising these public policies.

However, ethical and social values affect decisions made about the criteria (1) for identifying the range of risks, and of relevant empirical data needed for making judgments about them, that should be considered in public policy deliberations, and (2) for determining how well claims concerning risks should be supported by the available data in order to warrant that they have a decisive role in the deliberations. Consider the case of public policies about using GMOs. Concerning the range of data: is it sufficient for risk assessment only to be informed by data relevant to investigating the risks of using GMOs that may be occasioned by way of physical/chemical/biological mechanisms directly triggered by events within their modified genomes? Or: should data pertaining to the full range of ecological and socioeconomic effects of using them, in the environments in which they are used and under the socioeconomic conditions of their use, also inform this assessment? Those interested in producing and using GMOs, in the light of their adhering to values of capital and the market, are likely to give a positive answer to the first question; those holding competing values, e.g., connected with respect for human rights and environmental sustainability, to the second. And, concerning the degree of support: the former – citing the ethical gravity of losses (both economic and, allegedly, for food security) that would be incurred by failing to use GMOs on a wide scale – are likely to require less stringent standards of evidential appraisal than the latter.

Scientists, *qua* scientists, however, do not have special authority in the realm of values. Thus, their judgments, about the evidential support that claims about risks (and some other matters) have, may sometimes be reasonably (although not decisively) contested partly on value-laden grounds – as they have been in the GMO case, where the contestation has generated considerable controversy, and continues to do so. It follows that, in the context of deliberations about public policy, unless scientists engage with representatives of all stakeholders in the outcomes of the policies (as, for the most part, has not happened in the GMO case) – taking into account that their competing values may lead to making different decisions about what are the relevant data, as well as about the degree of support required for their claims about risks to gain the required credibility to inform the deliberations; and respecting "tempered equality" of participants in the dialogue (Longino) – their trustworthiness is put into question and their authority diminished.

1.

In a letter, dated June 29th, 2016, 135 Nobel laureates made the following claims, among others,[ii] related to using GMOs (genetically modified organisms) in agriculture:

(i) "Scientific and regulatory agencies around the world have repeatedly and consistently found crops and foods improved through biotechnology to be as safe as, if not safer than those derived from any other method of production."

(ii) "There has never been a single confirmed case of a negative health outcome for humans or animals from their consumption."

(iii) "Their environmental impacts have been shown repeatedly to be less damaging to the environment, and a boon to global biodiversity" (Laureates Letter, 2016).

Reflecting the authority and esteem that tends to be accorded to Nobel laureates, the declaration was widely reported and taken to bolster the allegation that there is a *scientific consensus* that cultivating and harvesting genetically engineered crops, and consuming their products, is safe.[iii] The scientists who signed it aimed to assure the public that the three claims are well con-

firmed, and that public policy and regulatory deliberations should reflect them. The claims do not derive from outcomes of the research conducted by these scientists, for at most one or two of them (so far as I can tell, none) have themselves engaged in biosafety research. They were putting their authority behind the research and judgments of others, whom presumably they trusted. Even so, one might reasonable assume that they had, before signing the declaration, examined the relevant research and concurred with its outcomes, and had found good reason to tell us, as they do, (presumably based on a thorough examination of its writings and actions) that the opposition is "based on emotion and dogma contradicted by data" and that it "must be stopped." At the end of the paper, I will argue that the declaration misuses scientific authority and contributes to doubts about the trustworthiness of leading scientific authorities. My larger purpose, however, is to suggest **some** necessary conditions for re-establishing trust in scientific communities – bridging the gap between scientists and the public, and (the concern of de Martín-Melo & Intemann, 2018) – so that both the authority and integrity of science, and the conditions for strengthening democratic societies, are enhanced

2.

First, some more general remarks. I maintain that the deliberations out of which arise public polices having to do with introducing, using and regulating technoscientific innovations (I only have time to discuss GEOs) should consider:

(1) questions about the *efficacy* of the proposed uses are addressed – and about their *safety*, specifically about how well available empirical evidence confirms that the proposed uses do not occasion harmful effects (or risks of causing harmful effects);

(2) the values reflected in the interests expected to be served by the proposed uses, as well as questions about whether interests expected to be served by competing values may be disadvantaged by them, and priorities among the competing interests;

(3) identified potential alternatives to using these innovations – including fundamentally different kinds of practices – as well as how using them compares to the proposed uses with respect to efficacy and safety (and other potential benefits).[iv]

Of these conditions only (1) is uncontroversial and generally followed (although there are disagreements about how it ought to be followed) in public policy deliberations.[v] Clearly satisfactory answers to the questions about efficacy and safety depend on trustworthy and reliable scientific input. I will not question that scientific research has reliably established the efficacy of the GEOs that have already been approved by regulatory bodies for agricultural use, for the most part GEOs with herbicide-resistant and insecticidal properties.[vi] Efficacy does not imply safety, however, and the research approaches (in molecular biology, biotechnology, etc) within which efficacy is established do not suffice for engaging in research dealing with safety. However, many regulatory practices presuppose that scientific input, pertaining to deliberations about safety – like that about efficacy – is obtained prior to consideration of (2) and (3), and to entanglement with value questions. Hence, the currency of the terms "scientific risk assessments" and

"scientific safety studies", areas of research in which scientific/technical "experts" should be granted authority.

One needs to be wary here, for "safe" and "risk" are 'thick ethical terms'. Scientific safety studies cannot be fully separate from entanglement with values and obligations. Thus, e.g. (simplifying a little), 'using X *is* unsafe' implies (*ceteris paribus*) 'X *should* not be used, unless appropriate precautions are taken.' And, when scientists conclude, on the basis of their investigations, that 'using X is safe', they intend it to follow (and to have impact at step (2)), that *ceteris paribus* 'it is improper to impede using X'.[vii] This does not mean that, in the course of empirical research in scientific safety studies, value-laden terms are used in articulating hypotheses and reporting empirical data. The link between the results of the empirical research and the subsequent value judgments depends on a step (call it step (0)), casually made prior to the empirical investigations. At step (0), the set of possible unintended collateral effects of using X is scrutinized, and those that are identified as harmful (as risks)[viii] – obviously value judgments are made here – are then investigated for such matters as the probability and magnitude of their possible occurrence, and its being countered by introducing scientifically informed regulations. In the investigation, the possible collateral effects are characterized, not with thick ethical terms, but with theoretical and observational terms deployed in relevant scientific fields, like molecular biology, chemistry, soil sciences and physiology (whose terms have no value connotations). Then, 'using X is safe' may be concluded,[ix] – usually qualified by 'provided that it is used in accordance with stipulated regulations' – if the investigations confirm that none of the investigated effects would occur with significant magnitude and probability when X is used in accordance with the regulations. This account is consistent with the picture of scientific safety studies that has step (1) preceding steps (2) and (3); but it clarifies that the move from empirically confirmed results at (1) to the claim the value-implicated 'X is safe' and to value judgments of relevance at (2) rests upon value judgments made at step (0). It follows that the conclusion, 'X is safe', might appropriately be challenged – without thereby challenging the scientists' judgments about each of the particular possible effects investigated – on the basis of the value judgment that not all the harmful possible effects of using X were identified at (0).

The outcomes of "scientific" safety studies usually constitute the only input to the deliberations of the 'technical' commissions that participate in public policy deliberations about using and regulating technoscientific objects. In these studies (in the GEO case), at step (0), the possible effects identified as harmful are a subset of those that may be occasioned by way of physical/chemical/biological mechanisms directly triggered by events within the modified genomes of plants. One can identify *two ways in which the adequacy of these studies might be challenged*.[x]

First: Conclusions drawn about the safety of using V (a genetically engineered plant variety) could be challenged on the ground that the subset chosen for investigation does not include some possible effects, with similar mechanisms, that are of of special salience for those who uphold a particular value-outlook.[xi] For them, even well conducted studies on the the items of the subset chosen will be insufficient to confirm that using V is safe.[xii] Challenges of this type can be

resolved (in principle) by conducting more scientific studies of the same kind after having identified a larger relevant subset.[xiii]

Second: Their adequacy could be challenged by those, who object that the set from which the subsets are chosen for "scientific safety studies" is not sufficiently encompassing. For them, deliberations about the safety of using GE-plants should be informed by appropriate empirical investigations, not only of potential effects occasioned by way of physical/chemical/biological mechanisms directly triggered by events within their modified genomes, but also the full range of potential ecological and socioeconomic effects occasioned by using them in the environments (agroecosystems) of their actual or intended use, and under the socioeconomic conditions of their use, taking fully into account that the potential effects vary from variety to variety and species to plant species. Upholding values of respect for human rights, democratic participation and environmental sustainability, which are opposed to those of capital and the market, often motivates challenges of this kind. These potential effects cannot *all* be investigated in "scientific safety studies," for they require utilizing ecological, human and social categories that have no place in research in such areas as physics, chemistry, and molecular biology, and that may include thick ethical terms (e.g., food security, being poisoned).[xiv] To investigate them empirically, therefore, requires adopting methodological approaches that are not reducible to those used in the indicated scientific areas, and that are generally outside of the expertise of scientists trained in the methodologies appropriate to them. The expertise required to engage in research that leads to the development of GEOs is quite different from that required for studies about the safety of using them.

At issue here are not only concerns about risks (potential harmful effects). Farmers (and their communities) in many areas of the world have suffered serious health problems because of having been exposed to glyphosate (the principal active ingredient in the widely used herbicide, RoundUp) sprayed on fields planted with glyphosate-resistant GEOs.[xv] They are unimpressed when told that the varieties of GEOs planted in these fields had undergone and passed "scientific safety tests." They know from their experience (even if it is not well recorded in peer reviewed studies) that, regardless of what was the case in the conditions of the tests, it is not safe to cultivate these GEOs (which require the accompanying use of glyphosate) in the ways and under the conditions in which they are used in their locales. And, they continue to be unimpressed when the manufactures and regulators of the GEOs insist that the problem was not with cultivating the GEOs, but with using glyphosate without heed to stipulated regulations for safe use,[xvi] for they have good reason to believe that the sellers of GEOs and glyphosate know that they will in fact not be used in accordance with these regulations.[xvii]

3.

Summing up, ethical and social values properly affect decisions (at step 0)) made about the criteria to be deployed for identifying the range of risks that should be considered in public policy deliberations, and of the relevant kinds of empirical data needed for making judgments about them. They also – consistent with maintaining that judgments about safety (step (1)) can be settled

prior to steps (2) and (3) – also affect the standards deployed for determining how well claims about risks should be supported (by the available empirical data) – in order to ensure that risks are dealt with properly in public policy deliberations.

Those who uphold values of capital and the market (agribusiness corporations, governments that prioritize economic growth, etc) are likely to cite the ethical gravity of losses (both economic and, allegedly, for food security) that would be incurred by failing to use GEOs on a wide scale; and consequently to require less stringent standards of evidential appraisal than those who uphold values of respect for human rights, democratic participation and environmental sustainability, who are likely to adopt precautionary stances that permit time for research incorporating more stringent standards to be met.[xviii] Similarly, those who uphold the latter values are likely to emphasize the importance of step (3): investigating alternatives to the food/agricultural system, in which using GEOs and the use of agrotoxics are acquiring ever larger roles, alternatives such as agroecology, a scientifically-informed approach to agriculture that attends simultaneously to production, sustainability, social health, strengthening the values and cultures of local communities, and to furthering the practices needed to implement policies of food sovereignty – and to urge the public support of research, in which are adopted strategies appropriate for dealing with the human, ecological and social dimensions of agroecosystems.[xix]

Scientists, *qua* scientists, however, do not have authority in the realm of ethical and social values. The values they uphold, even when widely shared, do not trump those upheld by other groups in democratic public policy deliberations. Thus, their judgments, about the evidential support that claims about the safety of planting GEO crops and consuming their products have, may sometimes be reasonably contested partly on value-laden grounds (cf. de Melo-Martín & Intemann, 2017, p. 131). That contestation cannot be rebutted by appeal to the alleged "scientific consensus" that GEOs (or, particular varieties of them) are safe. Apart from the fact that actually there is no such consensus, manifestly so among experts in biosafety investigations,[xx] if there were, it would likely secrete the scientists' shared value commitments, a matter on which they have no authority. Appeal to such an alleged consensus covers up the role of upholding the values of capital and the market in affirming it.

It follows that, in the context of deliberations about public policy, the trustworthiness of scientists is put into question and their authority unmerited,

- unless they engage with representatives of all stakeholders in the outcomes of the policies (as, for the most part, has not happened in the GEO case);
- unless, moreover, in doing so – respecting what Longino (2002, p. 129–135) calls "tempered equality" of participants in the deliberations – , they take into account that upholding competing values (e.g., of company-employed scientists and family farmers) may lead to making different judgments concerning relevant data, hypotheses to investigate, and approaches to farming, as well as concerning the degree of support required for claims about safety to merit credibility.

4.

5

Let us now return to the three claims (introduced at the outset) that the 135 Nobel laureates endorsed:[xxi]

These claims are ambiguous, misleading, in some instances false, and apparently made without acquaintance with the relevant studies and arguments of their critics. (i) is false: I am not aware of any agency that has compared the safety of GEO crops and their food products with that of agroecological (or organic farming) methods of production – the agencies have not sought out the results of research dealing with that comparison (and very little of it has been conducted). At most, they have found GEO crops and products to be at least as safe as conventional high-input crops and their products, but that doesn't respond to the critics who endorse agroecological methods of production. (ii) is probably true – but misleading: it does not mention that epidemiological studies of consumption of GEOs have not been conducted,[xxii] to a large extent because legal prohibition of labelling GEO products poses probably an insurmountable impediment to conducting them; and that it is well documented that cultivating GEOs has occasioned health problems for numerous farmers who have been exposed to the agrotoxics, whose use is integral to the cultivation of certain varieties of GEOs. (iii) is ambiguous: the environmental impacts may indeed be less damaging than those of conventional high-input agriculture; but they are incomparably more damaging to the environment than agroecological farming that has environmental sustainability built into its fundamental objectives.

By dismissing criticisms like these "based on emotion and dogma contradicted by data," and not attempting to rebut them in a context where something like Longino's conditions are in place, the scientists undermine the authority that science should be able to demand to be recognized; and they weaken the contribution that science could make to democratic policy deliberations.

References

Bombardi, Larissa M. (2017) Geografia do Uso de Agrotóxicos no Brasil e Conexões com a União Europeia. E-book, <https://drive.google.com/file/d/1ci7nzJPm_J6XYNkdv_rt-nbFmOETH80G/view>. São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

De Melo-Martín, I. and Intemann, K. (2018) *The Fight against Doubt: How to bridge the gap between scientists and the Public.* New York: Oxford University Press.

Hilbeck, A., Binimelis, R., Defarge, N., Steinbrecher, R., Székács, A., Wickson, F., Antoniou, M., Bereano, P. L., Clark, E. A., Hansen, M., Novotny, E., Heinemann, J., Meyer, H., Shiva, V. & Wynne, B. (2015) No scientific consensus on GEO safety. *Environmental Sciences Europe* 27: 4–9.

Human Rights Watch (2018) The Failing Response to pesticide Drift in Brazil's Rural Communities, July 20, 2018, <https://www.hrw.org/report/2018/07/20/you-dont-want-breathe-poison-anymore/failing-response-pesticide-drift-brazils>.

Krimsky, S. (2015) An illusory consensus behind GEO health assessment. *Science, Technology and Human Values* 40 (6): 883–914.

Lacey, H. (2005) *Values and Objectivity in Science; current controversy about transgenic crops.* Lanham, MD: Lexington Books.

— (2015a) Food and agricultural systems for the future: science, emancipation and human flourishing. *Journal of Critical Realism* 14 (3), 2015: 272–286.

— (2015b) Agroécologie : la science et les valeurs de la justice sociale, de la démocratie et de la durabilité. *Ecologie et Politique*, No. 51, 2015: 27–40.

— (2016) Science, respect for nature, and human well-being: democratic values and the responsibilities of scientists today. *Foundations of Science* 21(1): 883–914.

— (2017) The safety of using genetically engineered organism: empirical evidence and value judgments. *Public Affairs Quarterly* 31 (4): 259–279.

Lacey, H., Corrêa Leite, J., Oliveira, M.B., & Mariconda, P.r. (2015a) Transgênicos: malefícios, invasões e diálogo. *JC Notícias*, Edition 5167 (April 30, /2015), http://www.jornaldaciencia.org.br/edicoes/?url=http://jcnoticias.jornaldaciencia.org.br/9-transgenicos-maleficios-invasoes-e-dialogo/.

— (2015b) Transgênicos: diálogo. *JC Notícias*, Edition 5182 (May 22/2015), http://www.jornaldaciencia.org.br/edicoes/?url=http://jcnoticias.jornaldaciencia.org.br/27-transgenicos-dialogo/.

Longino, H. (2002) *The Fate of Knowledge*. Princeton: Princeton University Press.

Laureates Letter (2016) "Laureates letter supporting precision agriculture," http://supportprecisionagriculture.org/nobel-laureate-gmo-letter_rjr.html .

US National Academies of Science, Engineering and Medicine (2017). *Genetically Engineered Crops: Experiences and Prospects*. Washington: National Academies Press.

Paganelli, A., Gnazzo, V, Acosta, H., López, S.L. & Carrasco, A.E. (2010) 'Glyphosate-based herbicides produce teratogenic effects on vertebrates by impairing retinoic acid signaling'. *Chemical Research in Toxicology* 23: 1586–1595.

Traavik, T. & Ching, L.L. (2007) *Biosafety first: Holistic approaches to risk and uncertainty in genetic engineering and genetically modified organisms*. Trondheim, Norway: Tapir Academic Press.

Appendix

The central concern of the letter signed by the Nobel laureates is to support the program of research on Golden Rice [a variety of genetically engineered rice] and to denounce opposition to it, especially that of the NGO, Greenpeace. In a longer work, I would also discuss critically the way in which the letter misleads both about the state of research on Golden Rice and about that character of criticisms that question the importance of this research.

(a) The letter states that Greenpeace "has spearheaded opposition to Golden Rice, which has the potential to reduce or eliminate much of the death and disease caused by a vitamin A deficiency, which has the greatest impact on the poorest people in Africa and Southeast Asia". It called upon "governments of the world to reject Greenpeace's campaign against Golden Rice specifically, and crops and foods improved through biotechnology in general; and to do everything in their power to oppose Greenpeace's actions and accelerate the access of farmers to all the tools of modern biology, especially seeds improved through biotechnology"; and concluded with the warning: "Opposition based on emotion and dogma contradicted by data must be stopped," accompanied by the rhetorical question: "How many poor people in the world must die before we consider this a 'crime against humanity'?"

(b) Around the same time, the US National Academies of Science, Engineering and Medicine (2017) pointed out that the International Rice Research Institute (IRRI) had stated reported: "Golden Rice will only be made available broadly to farmers and consumers if it is successfully developed into rice varieties suitable for Asia, approved by national regulators, and shown to improve vitamin A status in community conditions. If Golden Rice is found to be safe and efficacious, a sustainable delivery program will ensure that Golden Rice is acceptable and accessible to those most in need" (p. 228). As of July 2016, IRRI was continuing research on developing varieties of Golden Rice for use in SE Asia, and (according to it) none of the conditions it stated had yet been met - it is for this reason that Golden Rice has not been introduced.

(c) Two years later, earlier this year (2018), IRRI asked the USFDA for an opinion regarding the safety of a variety of Golden Rice (called GR2E - the only variety yet submitted for regulatory approval - but not yet approved in any Asian country). FDA (May 24, 2018) endorsed the evaluation of IRRI (and the Australian regulatory body) that GR2E is safe for consumption, while pointing out that it is not intended for food or animal uses in USA. However, it added: "the concentration Beta-carotene in GR2E rice is too low to warrant a nutrient content claim." GR2E is safe but not nutritionally relevant.

(d) The signers of the letter, thus, were remarkably uninformed about the state of research on Golden Rice – and also about the views and stances of Greenpeace (I am not associated with Greenpeace). On its website Greenpeace states that its objective is to "ensure the ability of Earth to nurture life in all its diversity." It fits into the body of critics of using GMOs, who maintain that the dominant food-agricultural system (in which using GEOs has become for the time being a fundamental component) cannot respond adequately to the food and nutrition needs of the world's impoverished peoples (and the right to food security for everyone), and that these needs can best be ameliorated by the programs of agroecology and food sovereignty (Lacey, 2015a; 2015b) – and that programs for developing GEOs (like Golden Rice) are taking resources away from developing effective and lasting solutions to death and disease caused by vitamin A deficiency. Greenpeace has a respected place among these critics (and its "direct actions" and contributions to legal challenges are often appreciated by them). Of course, it would be legitimate to rebut the critics with argument and evidence. One wonders why the laureates did not attempt to do so.

(e) The credibility of pronouncements made by scientists of outstanding achievement is weakened when they sign letters like this one, accompanied by inflated, emotionally charged rhetoric, that has has a slender basis in fact. It would be enhanced if they entered into the type of dialogue, advocated by Helen Longino, in which scientists would "listen to" the evidence provided by relevant parties, attempt to understand critics, and not tar them without a hearing. Science has an indispensable contribution to make in policy deliberations; but it is not the determiner of policy. Science will be enhanced, and its role in democratic societies consolidated, if it claims only to have authority where it is actually warranted.

Notes

i **DRAFT** (not for citation outside of the PSA meeting in Seattle) – October 15, 2018. The text is a draft of the presentation I'm planning to make. The notes contain details that will be incorporated into an eventual completed paper.

ii See Appendix.

iii E.g., Mark Lynas (Cornell Alliance for Science), *A plea to Greenpeace*, <http://www.marklynas.org/2016/06/a-plea-to-greenpeace/>.
In this paper I only consider GEOs used in agriculture. I take for granted that claims to the effect that using GEOs is safe refer to GEOs that have passed safety tests, including those currently available on the market. (Obviously an unsafe GEO could be developed. Some varieties of GEOs have been developed that, after failing to pass safety tests, were not released for use.)

iv More fully developed and defended in Lacey (2005), Part 2.

v Deliberations concerning (2) and (3) cannot be settled in scientific inquiry (sound empirical inquiry), but there are sound empirically-based inputs that are (or could be) relevant to them. The deliberations will not be satisfactory if they do not draw upon these inputs. (See Lacey, 2005.)

vi Claims about efficacy need to be stated in a more qualified and nuanced way. I also will not contest that the claim that scientific research has not provided compelling evidence that consuming GEO products is unsafe health-wise. (The absence of compelling evidence that GEO products are unsafe to consume does not mean that there is compelling evidence that they are safe to consume – it depends on whether or not the necessary research has been conducted.)

vii The *ceteris paribus* qualification is needed to take into account that sometimes considerations, not reducible to safety ones, may properly be appealed to.

viii I will not discuss here how this set is generated – e.g., from considering past investigations, role of values in it, stakeholders' concerns, etc) – and who (holding what values?) makes (and should make) the identification of what should be considered harmful? following what kinds of deliberations? and who should be represented in the deliberations?.

ix To conclude on the basis of empirical investigation that 'X is safe' requires showing one-by-one that each member of the set of anticipated effect (judged to be harmful) is unlikely to occur at sufficient magnitude under the conditions imposed by proposed regulations. This presupposes: (a) an inductive move to unanticipated effects; and (b) that representative cases of all the effects, that should be labelled potentially harmful, are members of the set.

x I have argued elsewhere that here methodological and value considerations mutually reinforce each other (Lacey, 2017). Proponents of using GEOs often say that these safety studies investigate the risks occasioned by the GEOs themselves, and not those occasioned by the accompaniments of using them in agroecosystems or by socioeconomic mechanisms.

xi E.g., effects on soil microorganisms, a matter especially salient for those who regard maintaining soil fertility as indispensable for sustainable agriculture.

xii The studies, which have produced many of the results that have actually informed public policy and regulatory decisions, have been criticized for having a number of kinds of shortcomings (e.g., connected with conflicts of interest, and the use of intellectual property rights to maintain studies secret and so unavailable for replication and independent confirmation). Value judgments pervade these criticisms and their rebuttals. I will not attend to the questions that arise here.

xiii Such challenges might be deemed irrelevant by those who reject the value-outlook for which the possible effects have special salience, and so who reject the need for the further studies. Those adhering to the values of capital and the market sometimes take such a stand. How reasonable that might be depends on the arguments offered against holding the value-outlook in question.

xiv For elaboration see Lacey (2016; 2017).

xv For documentation, see, e.g., Bombardi (2017); Paganelli, et al. (2010); Human Rights Watch (2018).

xvi After a jury in California recently ruled that Monsanto was responsible for a man's being afflicted with cancer, and imposed a huge fine on it because it – for it was deemed that Monsanto had "acted with malice" in not providing warning on its label of the risks to health occasioned by using Roundup – the President of Bayer (that has now incorporated Monsanto) responded: "The correct use of Roundup doesn't present a risk to health" (reference to be added). [Monsanto has appealed the ruling.]

xvii Three years ago, when representatives of farmers – who had been poisoned in this way – came to present their testimony at a meeting of the "technical" commission in Brazil (CTNBio) that had appraised a particular variety of GEOs as safe, they were not granted a hearing since (most members of the commission maintained) they were bearers only of anecdotal (not scientific) evidence that had no relevance to the conclusions of scientific safety studies. When they then disrupted the meeting (and others of their group prevented the planting of a new variety of GEOs by invading a nursery and pulling up all the seedlings), they were denounced by major scientific organizations as having no respect for science, and acting on the

basis of "emotion and dogma." For criticisms of this stance taken by the majority of members of CTNBio, and a response to a rebuttal of the criticism, see Lacey, et al. (2015a; 2015b), articles published in *JC Notícias*, a daily e-newsletter of *Jornal da Ciência*, a publication of SBPC (Brazilian Society for the Advancement of Science).

The narrow scope of "scientific safety studies" is sometimes justified on the ground that the investigations of the social impact of using GEOs is not "scientific," for the methodologies adopted in them are not reducible to those adopted in the mainstream areas of science mentioned above. Be that as it may: I won't quibble about how to use the term "scientific" (a thick ethical term); the investigations in question are (when properly conducted) systematic empirical investigations. If they don't count as "scientific", that would imply that the results of "scientific" investigations cannot provide sufficient input into deliberations concerning public policies about safety, and would need to be supplemented with input from other kinds of empirical investigations.

[xviii] See Lacey (2017).

[xix] For details, see Lacey (2005; 2015a; 2015b).

[xx] See, e.g., Hilbeck, et al. (2015); Krimsky (2015); Traavik & Ching (2007).

[xxi] See Appendix.

[xxii] Unless all the relevant research has been conducted (and it has not been in this case), the absence of compelling evidence that GEO products are unsafe to consume does not imply that there is compelling evidence that they are safe to consume – and it has nothing to do with harms that may be caused by, e.g., contact with an agrotoxic, rather than by consumption.

The Reference Class Problem for Credit Valuation in Science

Carole J. Lee (c3@uw.edu)

Abstract: Scholars belong to multiple communities of credit simultaneously. When these communities disagree about how much credit to assign to a scholarly achievement, this raises a puzzle for decision theory models of credit-seeking in science. The reference class problem for credit valuation in science is the problem of determining to which of an agent's communities – which reference class – credit determinations should be indexed for any given act under any given state of nature. I will identify strategies and desiderata for resolving ambiguity in credit valuation due to this problem and explain how pursuing its solution could, ironically, lead to its dissolution.

1. Introduction

Within the scientific community, there is a common understanding that its reward system drives problematic behavior linked to publication patterns, pipeline retention, hypercompetitive scientific cultures, and reproducibility. Conversely, there is also a shared sentiment that, in order to change these cultures and behaviors in ways that would improve science, the scientific community must coordinate across institutions to change how credit is assigned at the level of the individual scientist (Alberts et al. 2014, Nosek et al. 2015, Aalbersberg et al. 2017, National Academies of Sciences 2018, National Science Foundation 2015, Blank et al. 2017). The hope is

that increasing individual researchers' incentives towards increased transparency and openness will improve the integrity, reproducibility, and accuracy of the published record.[1]

Analogously, philosophers working in the "credit economy" tradition adopt the working assumption that there is some amount of credit that agents can accrue for different acts under different states of nature. This assumption allows them to use decision theory to model how credit-seeking among individual scientists can give rise to behavior and norms that support or thwart the achievement of community-wide goals. When, in the aggregate, individual credit-seeking cuts against collective ends, their approach can explore how changes to individuals' incentive structures can nudge and redirect individual behavior (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Bright 2017, Heesen 2017, Kitcher 1990, Strevens 2003, Zollman 2018). Different philosophers make different assumptions about the norms by which credit gets allotted – for example, whether credit is best thought of as all-or-nothing (Strevens 2003, Bright 2017, Heesen 2017) or as something that may come in degrees (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Zollman 2018). However, the general approach assumes that there is some precise way to assign credit to different acts under different states of nature – an assumption that allows these philosophers to model credit-seeking behavior and the emergence of scientific norms in formally tractable ways.

But, how much credit gets assigned to any given act under any given state of nature? Just as each of us simultaneously belongs to multiple social categories each of which is tied to implied social hierarchies (Macrae, Bodenhausen, and Milne 1995, Crenshaw 1989), each

[1] Institutions can also experience incentives that promote or thwart scientific ends (Lee and Moher 2017).

scholar simultaneously belongs to multiple communities of value with implied social hierarchies for assigning credit. To which of an agent's communities – which reference class – should credit determinations be indexed and why?

In this paper, I will use examples from the current context of science's complex and dynamic culture to motivate and illuminate what I will call the *reference class problem for credit valuation in science*. I will identify a few strategies and desiderata for solving ambiguity in credit assignments due to the reference class problem. And, I will say a bit about how developing the resources needed to solve it could ultimately sow the seeds for its own dissolution.

2. The Reference Class Problem for Credit Valuation in Science

The contours of this puzzle about the "coin of recognition" (Merton 1968, 56) become visible when one moves beyond thinking about credit in generic, abstractions of scientific communities towards the heterogeneous communities we find today. I start from this slightly more concrete perspective because prestige requires recognition *by individuals and forums* that are themselves valued by credit-seeking scholars (Zuckerman and Merton 1971, Lee 2013): credit worthiness in science is a function of the individuals and systems designed to assess, allocate, dispute, and enforce it. Although some aspects of Zuckerman and Merton's narrative about the origins of the normative structure of science have been contested by historians (Csiszar 2015, Biagioli 2002), we see the social dynamics Zuckerman and Merton proposed clearly at play in contemporary science. For example, Nature Publishing Group recently found that – for the 18,354 authors in science, engineering, and medicine surveyed – the reputation of a journal is the primary factor driving choices about where to submit their work, where reputation is

primarily determined by the journal's impact factor and whether it is "seen as the place to publish the best research" (Nature Publishing Group 2015). Factors associated with a journal's ability to archive and disseminate research – things like a journal's time from acceptance to publication, indexing services, or Open Access options – were much less important.[2]

Within academia, each of us simultaneously belongs to multiple communities of value. The reference class problem arises when these different communities of value disagree about the amount of credit an agent accrues for choosing some act under some state of nature. Although I take this problem is be general, for the sake of clarity and simplicity in presentation, I will focus my examples on communities that can be described as having a nesting structure: for example, individual scholars belong to specific sub-disciplines, which are nested within disciplines, which are nested within a more general population of scholars. A sub-population that is nested within a population can have a credit sub-culture whose valuations differ from that of the population, whose valuations can differ from that of the super-population. In these cases, changing how narrowly or broadly one draws the boundaries of an agent's community of valuation can change the amount of credit assigned to a scholarly accomplishment. This gives rise to the *reference class problem for credit valuation in science*: to which of the agent's communities – which reference class – should credit valuations be indexed when determining the amount of credit the agent accrues for different acts under different states of nature?

[2] I recognize that some decision theorists, especially those working outside of philosophy, may reject or remain agnostic about attributing mental states such as beliefs to agents (Okasha 2016). However, because I understand credit and credit-seeking as sociological phenomena involving status beliefs such as these, I am committed to attributing beliefs to agents.

There are many examples across academia where nesting community structures can give rise to paradoxes and pathologies in credit assignments. For example, scholars' individual sense of what counts as quality work – their individual credit assignments – may deviate from what is endorsed in a sub-discipline or discipline's status hierarchy (Correll et al. 2017, Centola, Willer, and Macy 2005, Willer, Kuwabara, and Macy 2009). A puzzle that has cachet in a sub-discipline may be of peripheral importance within that discipline: for example, a more accurate technique for measuring how temperature cools with elevation considered critical in mountain meteorology and mountain ecology (Mindner, Mote, and Lundquist 2010) may have less visibility, despite its relevance, to the larger discipline of hydrology (Livneh et al. 2013). A question or technique that is thought to have high impact across fields (e.g., machine learning) may have little prominence within some of those fields.

Hypothetically speaking, one could imagine differences in valuations giving rise to a *Simpson's paradox in credit valuation*. Simpson's paradox is a phenomenon whereby a trend that appears in a population reverses or disappears when it is disaggregated into sub-populations (Blyth 1972). For example, a classic study found that, when looking at aggregate graduate school admissions data at UC Berkeley, women were, on the whole, less likely than men to be accepted; however, when the data was disaggregated into admitting departments, women were more likely than men to be admitted (Bickel, Hammel, and O'Connell 1975). Analogously, a *Simpson's paradox in credit valuation in science* would occur in cases where a population-level preference for scholarly product *a* versus *b* reverses when the population is disaggregated into its component sub-populations. In Simpson's Paradox cases, thinking more carefully about the context of evaluation usually leads to using a reference class that is finer-grained than the population-level. However, it's not clear whether this would always be the case in evaluations of

scientific credit. Hypothetically speaking, consider a hypothetical scenario in which an interdisciplinary project is not preferred by the individual disciplines represented by its authors or content, but is preferred when those disciplines are aggregated together. And, imagine that this project gets published in a journal, valued by those disciplines, that seeks papers of interest *across and beyond disciplines* (not just within disciplines): this is one way to interpret, for example, *Science's* mission to publish papers that "merit recognition by the wider scientific community and general public. . . beyond that provided by specialty journals" (Science). Which reference class would be most relevant in evaluating the value of this project?

There are other ways of dividing scholarly communities into nesting structures that create tensions in credit assignments. The pressures a scholar may feel from the incentive structure impacting her department/school may be slightly different from the incentive structure impacting her university. A coarse but concrete way to see this is to think about the prestige structure reified and reinforced by ranking systems (Espeland and Sauder 2012, 2016, Sauder and Espeland 2006), which transform "the ways professional opportunities are distributed" (Espeland and Sauder 2016, 7). An untenured business school professor with a potentially high impact manuscript needs to burnish her prestige in the eyes of both her dean and her provost, since both will evaluate her tenure case. If her provost is working to gain stature on the Academic Rankings of World Universities [ARWU], the professor should submit her manuscript to *Science* or *Nature*, since the ARWU ranks universities by their publications in these journals (Academic Ranking of World Universities 2018). However, if her dean is trying to gain stature on the *Financial Times* International ranking of MBA programs, she should submit to one of the fifty business, economics, or psychology journals by which the FT ranking system evaluates Business

school prestige – notably, the journal list does not include *Science* or *Nature* (Ormans 2016). What should the business school professor do?

Finally, credit assignments can vary depending on how long a time window a scholar keeps in view. A coarse but concrete way to think about this is by looking at how metrics for evaluating scholarship change over time. Journal impact factors are becoming less useful measures for evaluating an individual's scholarly contribution: since the advent of the digital age, the most elite journals (including *Science* and *Nature*) are publishing a decreasing percentage of the top cited papers (Larivière, Lozano, and Gingras 2013); the relationship between journal impact factor and paper citations has declined over time (Lozano, Larivière, and Gingras 2012); and, the citation distributions between journals "overlap extensively" (Larivière et al. 2016). The current wisdom is that if quantitative indicators are to be used to evaluate research, it is more useful to use article-level metrics such as citations as well as alternative metrics such as downloads and views (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017). On the horizon, there are now calls for creating new metrics that can encourage researchers and journals to be transparent and open in their reporting practices (National Academies of Sciences 2018, Wilsdon et al. 2017, Aalbersberg et al. 2017). Note that, the rise of such metrics – as well as the growing meta-research literature that ranks journals by the replicability (Schimmack 2015) or sample size and statistical power of their published results (Fraley and Vazire 2014) – makes it possible for a journal's impact factor and epistemic credibility to come apart (Fang and Casadevall 2011).

Decision theorists capture the risky nature of individual choices by allowing for uncertainty about which states of the world will come to be; and, when the probabilities attached to different outcomes are understood subjectively, these models permit a kind of subjectivity in

estimates of expected credit for different acts. However, I hope the examples throughout this section animate genuine *ambiguity in credit* due to the reference class problem for credit valuation in science.

3. Strategies and Desiderata for Solving the Reference Class Problem

How might decision theorists try to solve the reference class problem for assigning credit in science? One possible approach argues for the "correctness" of using one community rather than another. For example, it might be tempting to argue that all prestige is discipline-based since many scholarly prizes are distributed for excellence in particular disciplines (e.g., Nobel prize, Fields prize, academic society prizes); and, even when research is funded or published in interdisciplinary contexts, it may be primarily evaluated on the basis of its disciplinary excellence (Lamont 2009, but see Lee et al. 2013). Indexing credit valuation to a particular community need not prevent scholars from outside that community from understanding the relative value of that contribution: for example, if one were to adopt the old-fashioned and problematic assumption that an article's impact can be measured by the impact factor of the journal in which it is published,[3] and one recognizes that citations rates vary across disciplines, one could use field-normalized percentiles to understand a paper's impact in a metric that is legible across fields (Hicks and Wouters 2015). Because this strategy for addressing the

[3] The citation distributions within journals are so skewed that it is statistically improper to infer the impact of an individual article on the basis of the impact factor of the journal in which it is published (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017, Larivière et al. 2016, Wilsdon et al. 2015).

reference class problem relies heavily on identifying the "right" community, defending the centrality of the chosen community as opposed to others is critical. For example, some may challenge the idea that disciplines should be the sole arbiter of credit: note that the awarding of some scientific prizes reach across disciplinary conceptions of excellence (e.g., consider winners of the MacArthur Genius Prize and the psychologists who have won the Nobel Prize in Economics).

Another possible approach creates an algorithm that calculates the credit value of a scholarly contribution by summing the credit valuation of multiple communities. This approach would need to identify exactly how much to weight each community's valuation – with a rationale for why – since different weightings could lead to different overall credit valuations.[4] Note that some scholars take this style of approach when trying to measure the relative prestige of journals: in particular, the Eigenfactor score rates journals according to the number of its incoming citations, where the "relative importance" of each incoming citation is contextualized by the frequency with which the citing journal is itself cited (West, Bergstrom, and Bergstrom 2010).

Those who may wish to model the implications of different approaches for solving the reference class problem may try to do so by setting up hypothetical communities that assign

[4] On the face of it, this may seem like a form of commensuration because it involves summing values to calculate an overall score (Espeland and Stevens 1998). However, the process of commensuration requires combining values across *qualitatively* different domains of value. For clearer examples of commensuration in scholarly evaluation, see Lee (2015).

community boundaries and credit assignments in *de facto* ways to see what kinds of behaviors and norms emerge.

However, to solve the underlying conceptual problem, one must provide theories of community and credit that address two fundamental but vexing questions. How should one define and gerrymander the boundaries of the relevant communities invoked in the proposed solution? And, how does one determine the amount of credit those communities would assign to different acts under different states of nature? These questions may not be independently answerable. The boundaries of a community may need to be defined in terms of patterns of shared lore among its members about how credit is accrued – shared beliefs that coordinate credit-seeking and enforcement behavior in cases where status beliefs are internalized as norms (Merton 1973) and in cases where they are not (Willer, Kuwabara, and Macy 2009, Ridgeway and Correll 2006). Conversely, in recognition that some community members can have more influence than others on the content of reigning status beliefs, a community's credit assignments may need to be defined with some reference to the causal patterns of interaction among specific individuals and clusters of individuals – including status judges who wield "social control through their evaluation of role-performance and their allocation of rewards for that performance" (Zuckerman and Merton 1971, 66). Note, however, that answers to these questions should not *exclusively* inform each other. Notably, we must be careful not allow the size of a scholarly population and/or the power of its status judges to fully determine the intellectual value of the questions pursued by any particular partition of the scholarly universe.

4. Conclusion

Scientific credit – the "coin of recognition" (Merton 1968, 56) – is assessed, allocated, disputed, and enforced by many different communities and institutions within science that support and sustain a multiplicity of status hierarchies. This gives rise to what I have called the reference class problem for credit valuation in science. Solving this problem requires developing rich theories of community and credit that are based on fine-grained information about the structure and status systems of complex scholarly networks. The irony of this assessment is that such investigation towards solving the reference class problem could ultimately sow the seeds for its own dissolution.

In particular, such study can render friable a critical assumption for both the reference class problem and for decision theory models: namely, that communities, once defined, assign determinate amounts of monistic credit for different acts under different states of nature – that credit "can vary quantitatively but not qualitatively" (Anderson 1993, xii).[5] Contrary to this, recent policy papers call for moving away from narrowly conceived measurements of research excellence towards broader ones that are sensitive to the diversity of individual researchers', programs', and academic institutions' research missions (Hicks and Wouters 2015, Wilsdon et al. 2015). Such work can include community-engaged scholarship that creates, disseminates, and implements knowledge in coordination with the public to identify social interventions, change social practice, and influence policy (Hicks and Wouters 2015, San Francisco Declaration on Research Assessment 2013, Boyer 1990, Escrigas et al. 2014). From the

[5] Note too that, for formal reasons, the assumption that individual credit assessments could be aggregated into a collective one is questionable given the challenges of combining individual preferences into collective ones (Arrow 1950).

perspective of these efforts, plurality in our notions of scholarly excellence and credit – and differences in valuation and prioritization practices between individuals and communities – may be best conceived, not as a logical problem to solve, but as a starting point for theorizing.

Acknowledgments: Many thanks to Christopher Adolph, Aileen Fyfe, Crystal Hall, Jessica Lundquist, Conor Mayo-Wilson, and Kevin Zollman for helpful conversations. This research used statistical consulting resources provided by the Center for Statistics and the Social Sciences, University of Washington.

References

Aalbersberg, IJsbrand Jan, Tom Appleyard, Sarah Brookhart, Todd Carpenter, Michael Clarke, Stephen Curry, Josh Dahl, Alex DeHaven, Eric Eich, Maryrose Franko, Len Freedman, Chris Graf, Sean Grant, Brooks Hanson, Heather Joseph, Véronique Kiermer, Bianca Kramer, Alan Kraut, Roshan Kumar Karn, Carole Lee, Aki MacFarlane, Maryann Martone, Evan Mayo-Wilson, Marcia McNutt, Meredith McPhail, David Mellor, David Moher, Alison Mudditt Mudditt, Brian Nosek, Belinda Orland, Tim Parker, Mark Parsons, Mark Patterson, Solange Santos, Carolyn Shore, Dan Simons, Bobbie Spellman, Jeff Spies, Matt Spitzer, Victoria Stodden, Sowmya Swaminathan, Deborah Sweet, Anne Tsui, and Simine Vazire. 2017. "Making science transparent by default; Introducing the TOP Statement." *OSF Preprints*. doi: https://doi.org/10.31219/osf.io/sm78t.

Academic Ranking of World Universities. 2018. "ShanghaiRanking's Academic Ranking of World Universities 2018 Press Release." accessed September 1.

http://www.shanghairanking.com/Academic-Ranking-of-World-Universities-2018-Press-Release.html.

Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. 2014. "Rescuing US biomedical research from its systematic flaws." *Proceedings of the National Academy of Sciences* 111 (16):5773-7.

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Arrow, Kenneth J. 1950. "A difficulty in the concept of social welfare." *Journal of Political Economy* 58 (4):328-46.

Biagioli, Mario. 2002. "From Book Censorship to Academic Peer Review." *Emergences: Journal for the Study of Media & Composite Cultures* 12 (1):11-45.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. "Sex bias in graduate admissions: Data from Berkeley." *Science* 187 (4175):398-404.

Blank, Rebecca, Ronald J. Daniels, Gary Gilliland, Amy Gutmann, Samuel Hawgood, Freeman A. Hrabowski, Martha E. Pollack, Vincent Price, L. Rafael Reif, and Mark S. Schlissel. 2017. "A new data effort to inform career choices in biomedicine." *Science* 358 (6369):1388-9.

Blyth, Colin R. 1972. "On Simpson's Paradox and the sure-thing principle." *Journal of the American Statistical Association* 67 (338):364-66.

Boyer, Ernest L. 1990. *Scholarship Reconsidered*. San Francisco, CA: The Carnegie Foundation for the Advancement of Teaching.

Bright, Liam Kofi. 2017. "On Fraud." *Philosophical Studies* 174:291-310.

Bruner, Justin, and Cailin O'Connor. 2017. "Power, Bargaining, and Collaboration." In *Scientific Collaboration and Collective Knowledge*, edited by Thomas Boyer-Kassem, Conor Mayo-Wilson and Michael Weisberg, 135-157. Oxford, UK: Oxford University Press.

Centola, Damon, Robb Willer, and Michael Macy. 2005. "The emporer's dilemma: A computational model of self-enforcing norms." *American Journal of Sociology* 110 (4):1009-40.

Correll, Shelley J., Cecilia L. Ridgeway, Ezra W. Zuckerman, Sharon Jank, Sara Jordan-Bloch, and Sandra Nakagawa. 2017. "It's the conventional thought that counts: How third-order inference produces status advantage." *American Sociological Review* 82 (2):297-327.

Crenshaw, Kimberle. 1989. "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics." *University of Chicago Legal Forum* 139:139-168.

Csiszar, Alex. 2015. "Objectivities in Print." In *Objectivity in Science: New Perspectives from Science and Technology Studies*, edited by Flavia Padovani, Alan Richardson and Jonathan Y. Tsou, 145-69. Cham, Switzerland: Springer International Publishing.

Escrigas, Cristina, Jesús Granados Sánchez, Budd Hall, and Rajesh Tandon. 2014. "Editor's introduction. Knowledge, engagement and higher education: Contributing to social change." In *Report: Higher Education in the World*, edited by Cristina Escrigas, Jesús Granados Sánchez, Budd Hall and Rajesh Tandon. Palgrave Macmillan.

Espeland, Wendy Nelson, and Michael Sauder. 2012. "The Dynamism of Indicators." In *Governance by Indicators: Global Power through Quantification and Rankings*, edited by Kevin Davis, Angelina Fisher, Benedict Kingsbury and Sally Engle Merry, 86-109. Oxford: Oxford University Press.

Espeland, Wendy Nelson, and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York, NY: Russell Sage Foundation.

Espeland, Wendy Nelson, and Mitchell L. Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24:313-43.

Fang, Ferric C., and Arturo Casadevall. 2011. "Retracted Science and the Retraction Index." *Infection and Immunity* 79 (10):3855-9.

Fraley, R. Chris, and Simine Vazire. 2014. "The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power." *PLOS ONE* 9 (10):e109019. doi: 10.1371/journal.pone.0109019.

Heesen, Remco. 2017. "Communism and the Incentive to Share in Science." *Philosophy of Science* 84:698-716.

Hicks, Diana, and Paul Wouters. 2015. "The Leiden manifesto for research metrics." *Nature* 520:429-31.

Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* LXXXVII (1):5-22.

Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Larivière, Vincent, Véronique Kiermar, Catriona J. MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. 2016. "A simple proposal for the publication of journal citation distributions." *BioRxiv*:062109.

Larivière, Vincent, George A. Lozano, and Yves Gingras. 2013. "Are elite journals declining?" *Journal of the Association for Information Science and Technology* 65 (4):649-55.

Lee, Carole J. 2013. "The limited effectiveness of prestige as an intervention on the health of medical journal publications." *Episteme* 10 (4):387-402.

Lee, Carole J. 2015. "Commensuration bias in peer review." *Philosophy of Science* 82:1272-83.

Lee, Carole J., and David Moher. 2017. "Promote Scientific Integrity via Journal Peer Review." *Science* 357 (6348):256-7.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in peer review." *Journal of the American Society for Information Science and Technology* 64 (1):2-17.

Livneh, Ben, Eric A. Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M. Andreadis, Edwin P. Maurer, and Dennis P. Lettenmaier. 2013. "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions." *Journal of Climate* 26 (23):9384-9392.

Lozano, George A., Vincent Larivière, and Yves Gingras. 2012. "The weakening relationship between the Impact Factor and papers' citations in the digital age." *Journal of the American Society for Information Science and Technology* 63 (11):2140-45.

Macrae, C. Neil, Galen V. Bodenhausen, and Alan B. Milne. 1995. "The Dissection of Selection in Person Perception: Inhibitory Processes in Social Stereotyping." *Journal of Personality and Social Psychology* 69 (3):397-407.

Merton, Robert K. 1968. "The matthew effect in science." *Science* 1968:56-63.

Merton, Robert K. 1973. "The normative structure of science." In *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Norman W. Storer, 267-78. Chicago, IL: University of Chicago Press.

Mindner, Justin R., Philip W. Mote, and Jessica D. Lundquist. 2010. "Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains." *Journal of Geophysical Research: Atmospheres* 115. doi: https://doi.org/10.1029/2009JD013493.

National Academies of Sciences, Engineering, and Medicine,. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, D.C.: The National Academies Press.

National Science Foundation. 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. In *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*.

Nature Publishing Group. 2015. "Author Insights 2015 Survey."

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Mahlhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. "Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility." *Science* 348 (6242):1422-5. doi: 10.1126/science.aab2374.

Okasha, Samir. 2016. "On the interpretation of decision theory." *Economics & Philosophy* 32 (3):409-33.

Ormans, Laurent. 2016. "50 Journals used in FT research." accessed September 1. https://www.ft.com/content/3405a512-5cbb-11e1-8f1f-00144feabdc0.

Ridgeway, Cecilia L., and Shelley J. Correll. 2006. "Consensus and the creation and status beliefs." *Social Forces* 85 (1):431-53.

Rubin, Hannah, and Cailin O'Connor. 2018. "Discrimination and Collaboration in Science." *Philosophy of Science* 85:380-402.

San Francisco Declaration on Research Assessment. 2013. "The San Francisco Declaration on Research Assessment (DORA)." accessed September 1. https://sfdora.org/read/.

Sauder, Michael, and Wendy Nelson Espeland. 2006. "Strength in numbers? The advantages of multiple rankings." *Indiana Law Journal* 81 (1):205-27.

Schimmack, Ulrich. 2015. "Replicability Ranking of 26 Psychology Journals." January 18. https://replicationindex.wordpress.com/2015/08/13/replicability-ranking-of-26-psychology-journals/.

Science. "Mission and Scope." accessed September 1. http://sciencemag.org/about/mission-and-scope.

Strevens, Michael. 2003. "The role of the priority rule in science." *Journal of Philosophy* 100 (2):55-79.

West, Jevin D., Theodore C. Bergstrom, and Carl T. Bergstrom. 2010. "The Eigenfactor Metrics™: A network approach to assessing scholarly journals." *College & Research Libraries* 71 (3):236-44.

Willer, Robb, Ko Kuwabara, and Michael W. Macy. 2009. "The False Enforcement of Unpopular Norms." *American Journal of Sociology* 115 (2):451-90.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. 2015. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.

Wilsdon, James, Judit Bar-Ilan, Robert Frodeman, Elisabeth Lex, Isabella Peters, and Paul Wouters. 2017. Next-generation metrics: Responsible metrics and evaluation for open science. Report of the European Commission Expert Group on Altmetrics. European Commission.

Zollman, Kevin J. S. 2018. "The Credit Economy and the Economic Rationality of Science." *The Journal of Philosophy* 115:5-33.

Zuckerman, Harriet, and Robert K. Merton. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System." *Minerva* 9 (1):66-100.

Pragmatism and the content of quantum mechanics

Peter J. Lewis

Draft – please don't quote

Abstract

Pragmatism about quantum mechanics provides an attractive approach to the question of what quantum mechanics says. However, the conclusions reached by pragmatists concerning the content of quantum mechanics cannot be squared with the way that physicists use quantum mechanics to describe physical systems. In particular, attention to actual use results in ascribing content to claims about physical systems over a much wider range of contexts than countenanced by recent pragmatists. The resulting account of the content of quantum mechanics is much closer to quantum logic, and threatens the pragmatist conclusion that quantum mechanics requires no supplementation.

1. Introduction

Quantum mechanics is, notoriously, a theory in need of interpretation. But there is very little agreement on what kind of interpretation it needs. That is, there is very little agreement concerning what the foundational problems of quantum mechanics *are*, and without such agreement, there is little hope for a consensus concerning what an acceptable solution to the problems might look like.

Here is a way to divide up the territory. We can distinguish between *descriptive* and *normative* questions concerning quantum mechanics. Descriptive questions concern what quantum mechanics *says*—the *content* of the theory, as expressed in textbooks and used in labs. Normative questions concern what quantum mechanics *should* say—and in particular, whether it should say something different from what it actually does say.

All parties to the debates over the foundations of quantum mechanics would agree, I think, that there is a legitimate descriptive question concerning the content of quantum mechanics. Even those philosophers and physicists who think that quantum mechanics wears its interpretation on its sleeve at least feel the need to correct the mistaken impressions of *other* philosophers and physicists concerning what quantum mechanics says. The normative question presupposes an answer to the descriptive one: some think quantum mechanics is just fine the way it is, others contend that it needs to be replaced or supplemented with something radically different, and in large part this difference in attitude depends on prior differences concerning the answer to the descriptive question.

As an illustration, consider a fairly standard narrative concerning the descriptive and normative questions. Descriptively speaking, quantum mechanics depends on a distinction between measurements and non-measurements: measurements follow one dynamical law, the collapse dynamics, and non-measurements follow a different dynamical law, the Schrödinger dynamics. Since these two dynamical processes are incompatible, a precise formulation of quantum mechanics requires a precise dividing line between measurements and non-measurements. Quantum mechanics nowhere provides such a thing—and indeed, it seems highly unlikely that a term like “measurement” could be given a physically precise definition. So

descriptively speaking, quantum mechanics is inadequate as a physical theory. On the basis of this measurement problem, Bell (2004, 213–231) recommends replacing quantum mechanics with either a pilot-wave theory or a spontaneous collapse theory. For similar reasons, Wallace (2012, 35) recommends replacing quantum mechanics with a many-worlds theory.[1]

But not everybody concurs. There are alternative narratives according to which quantum mechanics, descriptively speaking, is just fine as it is, and hence there is no normative pressure to supplement or replace it. One prominent version proceeds from the quantum logic of von Neumann (1936) and Putnam (1975) through to the quantum information theory of Bub (2016). According to this approach, quantum mechanics describes a non-classical event space—in terms of truth values, a non-Boolean algebra, and in terms of probability ascriptions, a non-simplex distribution. No-go theorems (arguably) show that it is impossible to construct a set of events obeying classical Boolean logic or classical Kolmogorov probability that reproduces the empirical predictions of quantum mechanics. The implication is that in quantum mechanics we have discovered something important about the fundamental event structure of the world. Seeking to replace or supplement quantum mechanics with a theory obeying classical logic and classical probability theory amounts to a quixotic attempt to impose a structure on the world that it manifestly does not have (Bub 2016, 222). The measurement problem, on this account, results from a mistaken demand for a dynamical explanation of the individual events in the quantum structure, when no such explanation is available (Bub 2016, 223)

[1] Wallace takes the many-worlds theory to be a precise statement of the content of quantum mechanics, rather than a replacement for it. I take up the question of whether the many-worlds structure is present in quantum mechanics as it stands in section 2.

This fundamental difference of opinion—between those who take the measurement problem seriously and those who regard it as a pseudo-problem—continues to divide the foundations of physics community today. Hence the descriptive question—the question of what quantum mechanics actually *says*—remains a pressing one. In this paper, I argue for a particular way of approaching the descriptive question. The methodology is the pragmatist one of Healey (2012; 2017) and Friederich (2015), but the answer to the descriptive question that results from following this methodology, I argue, differs in an important way from the answers that Healey and Friederich give. I conclude by assessing the consequences of this answer to the descriptive question for the normative question.

2. The descriptive question

So how should we approach the descriptive question? Consider a straightforward realist approach to the content of scientific theories. A theory, at least in physics, is typically expressed using a particular mathematical structure. The *state* of a physical system is generally identified with a mathematical entity that resides in a particular abstract space, and the *dynamics* of the theory tell us how that state evolves over time. So, for example, in many applications of classical mechanics, the state of a physical system can be represented by a set of vectors in a three-dimensional Euclidean space, and the dynamical laws of Newtonian mechanics tell us how the set of vectors evolves over time. The interpretation of the mathematics is fairly straightforward: the vectors represent the positions and momenta of point-like particles, and classical mechanics tells us how the properties of the particles change.

Such an approach can equally be applied to quantum mechanics (Albert 1996). According to quantum mechanics, the state of a physical system is identified with a complex-valued function defined on a configuration space—a space with three dimensions for each particle in the system. A dynamical law, the Schrödinger equation, tells us how this function, the wave-function, changes over time. Then by analogy with classical mechanics, the wave-function must be a representation of the physical properties of the quantum system as they change over time.

The continuity with classical mechanics in the above account is attractive, but there are surprising consequences. For an *N*-particle system, the wave-function is defined over a 3*N*-dimensional configuration space, and it cannot be represented without loss in a three-dimensional space. This has led some to conclude that a straightforward realist reading of quantum mechanics shows that the three-dimensionality of our physical world is illusory (Albert 1996). Furthermore, if we model a measurement using quantum mechanics, the wave-function ends up with components corresponding to each possible outcome of the measurement—not just one outcome, as is the case classically. This leads Everettians like Wallace (2012) to conclude that a straightforward realist reading of quantum mechanics shows that every possible outcome of a measurement actually occurs.

These conclusions might be right, but do they simply follow from close attention to the structure of quantum mechanics? There are reasons to be suspicious. As Healey (2017, 116) notes, conclusions of this kind depend on the assumption that the wave-function plays the same descriptive role in quantum mechanics as the position-momentum vectors play in classical mechanics. If this assumption is itself up for grabs in the interpretation of quantum

mechanics, then neither of these conclusions is warranted. But how do we adjudicate the question of whether the wave-function describes physical systems or whether it has some other, non-descriptive role? Is there a metaphysically neutral methodology that could be used to answer this question? Healey (2012; 2017) and Friederich (2015) think that there is.

3. Pragmatism

Consider an analogy. "Stealing is bad" has the same grammatical structure as "Cherries are red". But it is far from clear that both sentences should be taken as descriptive. In particular, badness, taken as a property of actions, seems like a queer kind of property, imperceptible and disconnected from the other properties of the action. Expressivists seek to dissolve the problem of the nature of badness by claiming that a sentence like "Stealing is bad" should be taken as expressive rather than descriptive—as expressing our attitude towards stealing. Pragmatists further coopt expressivism as a variety of pragmatism (Price 2011, 9). Pragmatists stress the variety of uses of language, noting that sentences with superficially similar form can be used in radically different ways. "Cherries are red" is used to describe a class of objects, whereas "Stealing is bad" is used to express our attitude towards a class of actions.

Pragmatism, then, enjoins us to pay close attention to how a sentence is *used* in order to find out what it means. Healey (2012; 2017) and Friederich (2015) each suggest that the pragmatist approach provides us with a metaphysically neutral methodology for probing the content of quantum mechanics. That is, we can look at how various quantum mechanical claims are used by physicists in order to determine what those claims mean. This strikes me as a welcome suggestion. In the rest of this section I present the conclusions of their pragmatist

inquiries; in the next, I consider whether the language use of physicists actually supports those conclusions.

Healey (2012) distinguishes between *quantum claims* and *non-quantum magnitude claims*. The former explicitly mention quantum states, quantum probabilities, or other novel elements of the theory of quantum mechanics. The latter are claims about the magnitude of a physical quantity that do *not* involve quantum states, quantum probabilities etc. In keeping with the pragmatist methodology, Healey bases this distinction on the way the two kinds of claims are used. Non-quantum magnitude claims are used in a straightforwardly descriptive way. But quantum claims are used in a different way: they are used, not to *describe* a system, but to *prescribe* a user's degrees of belief in various non-quantum magnitude claims.

As an example, Healey appeals to the Interference experiments of Juffmann et al. (2009), in which C_{60} molecules are passed through an array of slits and then deposited on a silicon surface. To derive quantum mechanical predictions for this experimental arrangement, quantum states are ascribed to C_{60} molecules. That is, quantum claims of the form "The molecule has state $|\psi\rangle$" are used, via the Born rule, to ascribe probabilities to claims concerning the various possible locations of the molecules on the silicon surface. These latter claims—of the form "The molecule is located in region R"—are non-quantum magnitude claims. The job of the non-quantum magnitude claims is to describe the physical system, but the job of the quantum claims is to prescribe degrees of belief in the non-quantum magnitude claims for an appropriately situated observer. In this respect Healey's approach is like the expressivist's in ethics: claims that have superficially similar grammatical forms have very different functions.

Another important strand in the pragmatist approach concerns the role of decoherence. After the C_{60} molecule hits the silicon surface, complicated interactions with the surface mean that the state of the molecule-environment system becomes approximately diagonal when written as a density matrix in the position basis. This in turn insures that the probabilities ascribed by the Born rule to various claims about the molecule's position closely obey the probability axioms. But before the molecule encounters the silicon surface, its state is a coherent superposition—a state that is not even approximately diagonal, and for which the Born rule does not ascribe probabilities to location claims that closely obey the probability axioms. For such a state, the Born rule does not prescribe appropriate degrees of belief in the non-quantum location claims, and so assertion of such claims prior to decoherence is not *licensed* by quantum mechanics. Decoherence, then provides a demarcation between situations in which it is appropriate to have a well-defined degree of belief in a non-quantum magnitude claim, and situations in which it is not.

The central finding of the Healey-Friederich pragmatist approach is that attention to the use of quantum mechanical language shows that claims about the quantum state of a system are not used to describe that system. Hence, we should not think of the wave-function as a representation of the physical properties of the quantum system as they change over time. This perspective has the advantage that the measurement problem does not arise: if the wave-function doesn't represent the system, then we don't have to worry that the dynamical laws for wave-function evolution are different for measurements and non-measurements. In fact, if the quantum state is prescriptive, then the difference between measurements and non-

measurements arises quite naturally: the results of measurements have a direct and obvious influence on what you should believe.

Hence the pragmatist approach provides a clear answer to the descriptive question: quantum mechanics, in itself, says *nothing* about the world. As Healey (2017, 12) puts it, "quantum theory has no physical ontology". Rather, quantum mechanics tells us what to believe about non-quantum ontology—about particles, or in the case of quantum field theory, about fields. Furthermore, this answer to the descriptive question suggests an answer to the normative question: since the measurement problem doesn't arise, there is no motivation for supplementing or replacing quantum mechanics with something else.

4. Actual use, counterfactual content

Thus far, I have said little about the evidence that backs up Healey's claims about how quantum claims and non-quantum magnitude claims are used. Indeed, direct evidence from the language use of physicists is likely to be unenlightening: that a claim is asserted in a given context provides no direct evidence concerning whether its content is descriptive or prescriptive.

To fill this gap, Healey appeals to an inferentialist account of the link between use and meaning derived from the work of Robert Brandom (2000): the meaning of a claim is identified with the set of material inferences it licenses. So by looking at the way a claim is used in licensing inferences, we can gain evidence about what it means. And here the distinction between prescriptive quantum claims and descriptive non-quantum magnitude claims seems to be well motivated. In the practice of physics, a claim about the quantum state of a system is

used to infer Born probabilities, and nothing more. If Born probabilities are taken to be rational degrees of belief, then the prescriptive content of a quantum claim exhausts its meaning.

A non-quantum magnitude claim, on the other hand, can license a wide variety of inferences. From the claim that a C_{60} molecule is located in a particular region of the silicon surface, we can infer that an electron microscope will produce an image of the molecule if directed at that region (Juffmann et al. 2009, 2). We can infer that if the silicon surface is left untouched for two weeks, the C_{60} molecule will remain in the same place (Juffmann et al. 2009, 2). Under suitable conditions, we can infer that the C_{60} molecule will emits photons; under different conditions, that it will act as a nucleation core for molecular growth (Juffmann et al. 2009, 3). In other words, the inferences licensed by the non-quantum magnitude claim support the interpretation that the meaning of the claim is descriptive rather than merely prescriptive.[2]

So there is a good case to be made, I think, that actual use supports the distinction between prescriptive quantum claims and descriptive non-quantum magnitude claims. But there is a further strand to the Healey-Friederich interpretation, namely that non-quantum magnitude claims are only licensed after decoherence. This claim, I think, does not stand up so well to scrutiny.

Consider C_{60} interference again. After the molecule has adhered to the silicon surface, the state of the molecule is decoherent, and the claim that the molecule has a particular

[2] There is a sense in which the meaning of *any* claim is prescriptive according to the inferentialist program: the claim about the location of the molecule licenses an inference to a certain *degree of belief* that the electron microscope will produce an image of it. But still, there is a reasonable distinction here: the quantum claim licenses inferences only via the Born rule, whereas the non-quantum magnitude claim licenses inferences via a huge variety of schema typical of small physical objects. The latter is just what it is for a claim to be descriptive.

location is licensed—that is, it is appropriate to associate a particular degree of belief with the claim, and if that degree of belief is high enough, it is appropriate to assert the claim. But before the molecule has adhered to the silicon surface, the state of the molecule is coherent, and no claim about the location of the molecule is licensed—it is not appropriate to associate a degree of belief with such a claim, or to assert it. Similar considerations apply to properties other than location.

This seems to fly in the face of actual use. For example, in the description of the C_{60} interference experiment, Juffmann et al. (2009, 2) assert that "all transmitted particles arrive with the same speed," and "about 110cm behind the source, the molecules encounter the first diffraction grating," apparently ascribing both speed and location to C_{60} molecules prior to decoherence. This doesn't seem to be an isolated incident: physicists routinely talk of preparing, selecting, spraying, shooting and trapping particles, ions and molecules, and this talk typically involves making claims about these objects prior to any eventual decoherence.

It is possible, of course, that this is just "loose talk", or an indirect way of making claims about the quantum state of the systems concerned. But given the frequency of such claims, and given the reliance of the pragmatist methodology on *use*, this seems like a shaky game to play. It would be better, all things considered, if such claims could be accommodated within the pragmatist interpretation, rather than explained away as anomalies.

But there are obvious barriers to licensing non-quantum magnitude claims prior to decoherence. As Friederich (2015, 79) notes, the Born rule is only "reliable" when applied to decoherent states, in the sense that only for such states are the numbers it produces guaranteed to closely obey the probability axioms. Given some reasonable assumptions about

rationality, it is plausible that numbers that do not closely obey the probability axioms could not be rational degrees of belief. Furthermore, Healey argues that asserting a non-quantum magnitude claim prior to decoherence is likely to be misleading. For example, suppose one asserts (with Juffmann et al.) that "about 110cm behind the source, the molecules encounter the first diffraction grating." One might infer from this that each molecule passes through exactly one slit in the grating, and hence that the presence of the other slits is irrelevant, and hence that there is no possibility of interference (Healey 2012, 745).

So the pragmatist approach seems to face a dilemma: either it fails to accommodate the actual language use of physicists, or it licenses misleading assertions and irrational degrees of belief. Isn't there another way? I think there is. Consider a mundane claim like "There is beer in the fridge." In typical contexts, an assertion of this claim licenses the inference that if you were to go to the fridge and open the door, you could take a beer and drink it. Of course, you might not actually do this; maybe you don't want a beer. That is, the inference here is a counterfactual one. A good deal of the inferential content of our assertions has this counterfactual character.

Now return to the quantum context. Consider again the claim that "about 110cm behind the source, the molecules encounter the first diffraction grating." What content could that claim have? If we broaden the notion of inferential content to include counterfactual inferences, then the content seems fairly clear: if we were to replace the first diffraction grating with a detector taking up the same region of space, then the Born rule would ascribe a degree of belief close to 1 to detecting the molecules.

How does the inclusion of counterfactual content avoid the barriers to licensing non-quantum magnitude claims prior to decoherence? Note that the counterfactual content of the claim about the molecules involves a counterfactual intervention on the system—a counterfactual measurement. The counterfactual measurement induces counterfactual decoherence. The Born probabilities are conditional on this intervention and the associated decoherence, so the Born probabilities for various position claims concerning the molecules are not, after all, unreliable, in the sense of violating the probability axioms.

Neither should there be any danger of being misled by an assertion that the C_{60} molecules encounter the grating, because the counterfactual conditions implicit in the content of that assertion are distinct from the conditions that actually obtain in the apparatus. That you *could* detect the molecules at the diffraction grating, given a different experimental arrangement, doesn't license the inference that there *is* no interference, given the actual experimental arrangement. Admittedly, though, this amounts to a weakening of the content of position claims from the classical case, as spelled out in the next section.

5. A happy convergence?

I have argued that non-quantum magnitude claims have assertible content in a far wider range of contexts than countenanced by Healey or Friederich. If there is some counterfactual intervention on a system that would produce decoherence in the basis defined by a given observable, then claims about the values of that observable have content. And since counterfactual interventions only have to be realizable in principle, this means that claims about the value of an observable for a system *generally* have content, whether or not the

system *actually* decoheres in the basis defined by that observable. This has the welcome consequence that the frequent assertions made by physicists about the properties of systems prior to decoherence are contentful.

A potential cost of such permissiveness about content is that the structure of this content is, in general, non-Boolean. Consider again a C_{60} molecule that is approaching the first diffraction grating, and consider an assertion of "The molecule passes through the leftmost slit". This assertion has content, on the proposed view, because in principle there is an intervention on the system that would produce decoherence in a basis defined by an observable that distinguishes which slit the molecule passes through. Still, assertion of the claim would not be appropriate, simply because there are many slits in the grating, so the Born rule ascribes it a low probability. The same goes for every other slit in the grating. Nevertheless, the assertion that "The molecule passes through the leftmost slit, or the second to the left, or…" is assertible, since the Born rule ascribes it a probability close to 1. The disjunction is assertible, but none of the disjuncts is assertible. Since assertibility is a surrogate for truth in the pragmatist context, this is equivalent to saying that the disjunction is true, but none of the disjuncts is true.

One might take this to be unacceptable on the pragmatist view—especially if you endorse an inferentialist pragmatism, as Healey does. From a disjunctive claim you can straightforwardly infer that at least one of the disjuncts is true. If the content of a claim is identified with the inferences that it licenses, then part of the meaning of the disjunctive claim about the C_{60} molecule is that some assertion of the form "The molecule went through slit *x*" is true. Hence my proposal about content threatens to violate the inferentialist account of

meaning. The pragmatist interpretation of Healey and Friederich avoids this problem by insisting that claims about systems have meaning only after suitable decoherence.

Of course, pragmatism is not necessarily tied to an inferentialist account of meaning. But even given inferentialism, there is arguably no real problem here. Physicists are *selective* in the inferences they draw: from the disjunctive claim, they don't infer that the C_{60} molecule goes through some particular slit, so they don't infer a lack of interference. But they do infer that the molecule will arrive at the silicon surface, that it might radiate a photon in flight, and so forth. That is, the inferences drawn by physicists from their claims about pre-decoherent systems suggest that the non-Boolean structure of those claims is already *built in* to the meanings associated with those claims and revealed in inference.

This suggests that close attention to the way non-quantum magnitude claims are actually used leads to a happy convergence between pragmatism and the quantum logical approach. Physicists assert claims about particles even when the state does not decohere, and such claims seem to be meaningful. But physicists are not inclined on that basis to draw all the inferences that a full Boolean structure to their claims would license. Quantum mechanics apparently weakens the meaning of many claims about pre-decoherent physical systems, but without rendering those claims meaningless.

6. The normative question

As a methodology for addressing the *descriptive* question of the content of quantum mechanics, the pragmatist approach seems entirely appropriate: look to the *use* of physicists to determine what the various claims involved in the theory mean. At the hands of Healey and

Friederich, this approach yields the important insight that while non-quantum magnitude claims are used to describe physical system, quantum claims are used to prescribe appropriate degrees of belief in non-quantum magnitude claims. But Healey and Friederich go further, in limiting the assertibility of non-quantum magnitude claims to contexts in which the quantum state is decoherent in the relevant basis. This, I have argued, cannot be squared with the actual use of such claims. I propose instead that non-quantum magnitude claims *generally* have well-defined content, understood in terms of a counterfactual intervention on the system. This change to the pragmatist approach means that it ends up looking a lot like the quantum logical approach that preceded it. Indeed, the pragmatist approach might be regarded as a *justification* for quantum logical claims concerning the content of quantum mechanics.

But where does all this leave the *normative* question concerning whether quantum mechanics is fine as it is, or whether it should be supplemented or replaced? Healey and Friederich argue that quantum mechanics is fine as it is: if quantum claims do not describe physical systems, then there can be no conflict between the way that quantum mechanics describes systems during measurements and the way it describes them during non-measurements. If there is no measurement problem, then there is no motivation to replace such a successful theory. If, as Healey (2017, 12) maintains, quantum theory "states no facts about physical objects or events," then there can be no requirement that we come up with an *explanation* of quantum facts and events.

However, I have suggested that quantum theory has more content than the pragmatists countenance. In one sense, I agree that quantum theory states no facts: a quantum claim, such as the attribution of a quantum state to a system, is not a description. But in another sense,

there are distinctive quantum facts, or at least facts with a distinctive quantum structure: non-quantum magnitude claims about pre-decoherent systems exhibit the non-Boolean structure characteristic of quantum mechanics. This is the sense in which quantum logic gets things right.

Notably, though, the proponents of quantum logic *also* often take the view that quantum logic dissolves the measurement problem (e.g. Putnam 1975, 186). But this dissolution is widely regarded to be a failure (e.g. Bacciagaluppi 2009, 65) Once one has admitted that the structure of true (i.e. assertible) claims for a quantum system is non-Boolean, the question of *how* the world manages to instantiate this structure becomes legitimate and pressing. A denial that any explanation is required looks suspiciously like instrumentalism. And since any answer to this question goes beyond quantum mechanics as it stands, the call for explanation involves a demand to supplement quantum mechanics, or to replace it with something more fundamental.

Of course, given the no-go theorems, the path to an explanation of the structure of quantum facts is by no means clear. But neither do the no-go theorems show that an explanation is *impossible* (Friederich 2015, 161).[3] If the foregoing is correct, then pragmatism is an excellent way to *expose* the foundational problems of quantum mechanics, but it is not a means to *dissolve* them.

References

[3] Interestingly, Friederich (2015, 161) suggests supplementing quantum mechanics with sharp values for all observables, even though this seems at odds with his therapeutic aim of dissolving the foundational problems of quantum mechanics rather than solving them (2015, 6).

Albert, David Z. (1996), "Elementary quantum metaphysics," in J. T. Cushing, A. Fine and S. Goldstein (eds.), *Bohmian Mechanics and Quantum Theory: An Appraisal*. Dordrecht: Springer, 277-284.

Bacciagaluppi, Guido (2009), "Is logic empirical?" in K. Engesser, D. M. Gabbay and D. Lehmann (eds.), *Handbook of Quantum Logic and Quantum Structures.* Amsterdam: North-Holland, 49-78.

Bell, J. S. (2004), *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press.

Brandom, R. (2000), *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Bub, Jeffrey (2016), *Bananaworld: Quantum Mechanics for Primates*. Oxford: Oxford University Press.

Friederich, Simon (2015), *Interpreting Quantum Theory: A Therapeutic Approach*. Basingstoke: Palgrave Macmillan.

Healey, Richard (2012), "Quantum theory: a pragmatist approach," *British Journal for the Philosophy of Science* 63: 729-771.

Healey, Richard (2017), *The Quantum Revolution in Philosophy*. Oxford: Oxford University Press.

Juffmann, T., Truppe, S., Geyer, P., Major, A. G., Deachapunya, S., Ulbricht, H., and Arndt, M. (2009), "Wave and particle in molecular interference lithography," *Physical Review Letters* 103: 263601.

Price, Huw (2011), *Naturalism Without Mirrors*. Oxford: Oxford University Press.

Putnam, Hilary (1975), "The logic of quantum mechanics," in *Mathematics, Matter and Method: Philosophical Papers Volume 1*. Cambridge: Cambridge University Press.

von Neumann, John (1932), *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer-Verlag.

Wallace, David (2012), *The Emergent Multiverse*. Oxford: Oxford University Press.

Tool Migration: A Framework for Analyzing Cross-disciplinary Use of Mathematical Constructs

Chia-Hua Lin
University of South Carolina / KLI

Abstract. Mathematical formalisms that are constructed for inquiry in one disciplinary context are sometimes applied to another, a phenomenon that I call 'tool migration.' Philosophers of science have addressed the advantages of using migrated tools. In this paper, I argue that tool migration can be epistemically risky. I then develop an analytic framework for better understanding the risks that are implicit in tool migration. My approach shows that viewing mathematical constructs as tools while also acknowledging their representational features allows for a balanced understanding of knowledge production that are aided by the research tools migrated across disciplinary boundaries.

Keywords: Cross-disciplinarity, tool migration, epistemic risks

1. Introduction

Mathematical formalisms that are constructed for scientific inquiry in one disciplinary (or sub-disciplinary) context are applied to another. Philosophers of science have started paying attention to this cross-disciplinary aspect of scientific practice. For instance, the discussion of 'model transfer' concerns a relatively small set of mathematical models that are applied in multiple disciplinary contexts. Humphreys (2004) proposes that models that are transferred to study phenomena of a different domain owe their versatility to the computational tractability they afford. In contrast, Knuuttila and Loettger (2014, 2016) suggest that in addition to tractability, versatile models also offer conceptual frameworks for theorization, which they label 'model templates.' However, these analyses do not deal with the risks inherent in this aspect of scientific practice. Consider the use and development of game theory in evolutionary biology as an example. In importing game theory, which was originally conceived to describe strategic interaction between rational agents typically studied by social scientists, evolutionary biologists may need to modify the theory in order to generate knowledge about presumably non-rational agents, at least in many cases. One can then assume that any changes to the theory--between its established applications in social sciences and its novel uses in evolutionary biology--require special attention so as to avoid misinterpreting an analysis.

Despite the advantages, there might be risks associated with using mathematical constructs across disciplines. In this paper, I ask: might there be patterns of transfer that may undermine the effectiveness of the imported mathematical formulation? What would these

patterns, if any, look like? This paper is an attempt to explore the conditions in which importing mathematical constructs may be epistemically risky. To begin, I develop a framework to systematically characterize the landscape of mathematical importations. The goal of such a framework is two-fold. Proximately, the framework captures characteristics of migration that the current terminology, such as 'model transfer' or 'importing/exporting,' fails to discern. Ultimately, with this additional discernibility, I suggest that one may start to explore and identify patterns of importation that may be subject to epistemic risks, such as misinterpretation of an outcome produced by using a imported mathematical construct.

In Section 2, I argue that one can view mathematical constructs in science in terms of 'research tools' and that transporting such tools across disciplines, which I call 'tool migration,' can in some cases be a disservice to science. Next, I classify tool migration based on two kinds of contextual details that bear significance to the effectiveness of the migrated research tool in a foreign context. In Section 3, I apply this approach to the use and development of game theory in evolutionary biology. Finally, in Section 4, I discuss in what ways this tool migration framework, which is essentially a typology of four types of tool migration, may help to characterize epistemically risky patterns of tool migration.

2. Theoretical Background

Although the notion of epistemic risks associated with migration of mathematical constructs has not been explicitly addressed, the idea of viewing mathematical constructs as research tools follows from the discussion on the ontology of scientific models. Ever since the shift of attention to scientific practice (e.g., Hacking 1983), there has been a growing literature in which models in science are viewed as entities *detachable* from theory and data (e.g., Morrison 1999; Morgan and Morrison 1999). One recent predecessor to my tool migration account is a pragmatic approach to scientific models put forth by Boon and Knuuttila (2008). In their paper, which uses examples from engineering, they argue that scientific models are better understood as 'epistemic tools' instead of as representations of some target systems in the world. Boon and Knuuttila's argument draws heavily on the epistemological roles of scientific models in relation to the scientists who use them. According to them, scientific models allow their users "to understand, predict, or optimize the behavior of devices or the properties of diverse materials" (2008, 687). Thus, for an ontological account of scientific models to be productive and realistic, as they argue, it should be sensitive to the relation between the models and the modelers, i.e., the tools and their users. An adequate evaluation of Boon and Knuuttila's argument will take us far afield, but my work will show that both the representational and the pragmatic aspects are indispensable to a better understanding of the epistemic risks in tool migration.

2.1 Viewing mathematical constructs as research tools

In general terms, any mathematical construct that is to be *used or operated* in an algorithmic manner, and the outcome of whose operation is to be *interpreted* in order to answer a research question, is an example of what I am calling a research tool. Let me first unpack the operational aspect of a research tool.

Let's assume that the proper use of any mathematical constructs employed in scientific research is expected to produce consistent results. To achieve this consistency, then, a well-defined procedure needs to accompany such a construct so that anyone who follows the procedure expects, and is expected, to obtain the same outcome given the same input. For instance, when performing a game-theoretic analysis, one goes through a sequence of steps, such as: (i) identify the players and the acts available to them, (ii) identify the payouts in every set of acts, (iii) find the 'Nash equilibria,' which refers to a set of acts, one for each player, in which no player could improve his or her payoff by unilaterally changing act. A similar algorithmic procedure can be seen when applying, say, Newton's law of gravitation:

$$F_{grav} = G\frac{m_1 m_2}{r^2}. \qquad (1.1)$$

For example, the sequence of steps to obtain the magnitude of the gravitational force, F_{grav}, between any two objects includes: (i) identify the mass of each object, (ii) identify the distance between them, (iii) complete the equation in which 'm_1' and 'm_2' refer to the masses of the two objects, 'r' the distance in between, and 'G' the gravitational constant. In these two examples, when the first two steps produce consistent input, the third step is expected to generate the same output.

Moreover, concerning the interpretational aspect of a research tool, the output of a series of symbol assignments and manipulations can be understood *only through the lens of some interpretation*. The Nash-equilibrium of a game is a meaningful 'solution' in virtue of the usual understanding of the game-theoretic formulation of a problem. Similarly, the meaning of the value obtained through completing the equation in (1.1) is derived from the usual interpretation of the quantities appearing in the equation and the theoretical context in which those quantities are defined.

Finally, assume that something can be viewed as a tool if it serves as a means to an end. In this case, then, mathematical constructs like game theory or mathematical formulas can be seen as research tools. In the case of applying a mathematical construct, the goal of performing a sequence of prescribed steps goes beyond merely completing the calculation and obtaining a result. Instead, the output is to be interpreted so that one may solve a problem, answer a research question, or gain knowledge about a subject-matter. Thus, a mathematical construct that prescribes algorithmic symbol manipulation can be seen as a research tool, assisting its users to meet an end. Manipulating symbols is a means to the end that was specified during the mathematical formulation of the research problem.

2.2 Epistemic risks of tool migration

Another predecessor to my account is Morgan's discussion of the re-situating of knowledge (2014). According to her, knowledge production is necessarily 'situated,' and consequently, applying a piece of knowledge outside its initial context requires effort - different contextual situations require different 're-situating' strategies. The term 're-situation' thus captures what scientists do in practice to transport locally generated knowledge across contexts. As she argues, to make an instance of scientific knowledge accessible outside its production site, one needs to establish inferential links between the production site and the destination site. However, she suggests, whether a re-situation of knowledge contributes to scientific progress depends on whether the transport secures some sort of inferential safety.

Building from Morgan's notion of the re-situation of knowledge, I argue that cross-disciplinary use of research tools is epistemically risky. Given the locality of scientific knowledge production, applying scientific knowledge outside its production site may come with epistemic risks. For example, between the production site and a destination site, there may be incongruent disciplinary characteristics (e.g., implicit theoretical assumptions) that fail to be captured by the inferential strategy, such that knowledge from the former cannot be transferred to the latter. Similarly, we can assume that the construction of a research tool is also *situated* in nature. Namely, a research tool is conceived to be operated and to extend our knowledge concerning a subject-matter *given a particular disciplinary context*. It follows that cross-disciplinary use of research tools is as epistemically risky as re-situating knowledge. That is, the epistemic reliability (i.e., general ability or tendency to produce knowledge) of some research tool in one disciplinary context does not necessarily carry over to another.

The concept of 'tool migration' captures both the 'situated-ness' of a research tool that was established in its native discipline and the effort it takes to 're-situate' the tool in a foreign discipline. Naturally, in the process of uprooting a research tool, significant contextual details—ranging from implicit expertise to important background assumptions—may be stripped away. Likewise, during re-situation, new features may be introduced to the tool so as to treat a different subject matter in a new disciplinary context. Together, due to the possibility of losing or gaining significant contextual details, or both, a cross-disciplinary tool migration risks undermining the effectiveness of the tool. These risks include, for example, misinterpretation of the research result or failure to produce genuine knowledge. Thus, it follows that tool migration can in some cases be a disservice to the production of knowledge.

Acknowledging these challenges, some have argued against the cross-disciplinary effort to integrate disciplinary knowledge (e.g., van der Steen 1993). Alternatively, one might try to overcome these challenges so long as the risks are better understood and managed. To understand the risks, I suggest that we first look at the patterns of tool migration. Among these patterns, we might find that some of them could be epistemically risky. Having established the

notions of research tools and risks involved with tool migration, I turn to the contextual details that are closely related to a tool's epistemic performance.

2.3 Contextual details of a research tool: the target profile and the usage profile

The construction of a research tool is necessarily situated within a context. In order to compare and contrast between the native (or established) context and the foreign context of a migrated tool, I single out two major types of details.

The first type concerns the assumptions about the entities that are studied by a subject-matter for which the tool is developed. For instance, game theory defines what it considers as a game, a player, or an act. For simplicity, I call *all* the assumptions that a tool makes about its target entities the tool's 'target profile.'

The second type considers *the ways* in which one interprets the output from applying a tool in his or her research. In a game-theoretic analysis, for example, by following an algorithmic procedure, one obtains a solution of a game in the form of a Nash equilibrium. Depending on the game that one was analyzing, the solution could be understood as an explanation of economic behavior, or a prediction about it, or it could be used to optimize an strategic interaction. For simplicity, I call *all* the ways in which a tool is intended to be used, e.g., describing, predicting, optimizing, or explaining its target phenomenon, the tool's 'usage profile.'

Together, as I demonstrate in Section 4, the 'target profile' and 'usage profile' allow one to detect patterns of changes in the contextual details between the established use and the novel use of a research tool. They are able to do this because these two profiles offer a coarse resolution; looking through the lens of the target profile and usage profile, one zooms out from particular cases of tool migration so as to detect patterns of cross-disciplinary transport. Further analyses of these patterns will then shed lights on their associated epistemic risks.

2.4 Four types of tool migration

With the two profiles of a research tool and the two contexts in which the tool is used, i.e., a novel use and an established use, one can distinguish four types of tool migration.

First, compared to its established use, when a novel use of a tool catalyzes changes in both target and usage profiles, the tool migration is transformative, and therefore I call it a ***tool-transformation***. Second, in contrast, when both target and usage profiles remain more or less intact after the migration, the tool's novel use is considerably similar to its previous applications. Thus, I call such a case ***tool-application***. Between these two extreme types, there are novel uses of a research tool that alter only one of the two profiles but not both. When a tool changes its target profile but not its usage profile, I call it a ***tool-transfer,*** and when a tool changes its usage profile but not the target profile, I call it a ***tool-adaptation***. See **Table 1** for a summary.

Table 1
A Typology of Tool Migration

Between established and novel uses of a research tool	**Usage profile remains**	**Usage profile deviates**
Target profile remains	'Tool-application'	'Tool-adaptation'
Target profile deviates	'Tool-transfer'	'Tool-transformation'

Among these four types of tool migration, tool-transfer is arguably the most familiar to the philosophers of science. Humphreys coins the term ‘computational templates’ to refer to a relatively small number of mathematical equations that are applied to investigate different domains of phenomena (2002, 2004). Bailer-Jones (2009) discusses such a scientific practice in terms of mathematical analogy. For one example, Newton's law of gravitation was intentionally sought after to model electrostatic force (see Bailer-Jones 2009 for a detailed account). The important parallel between the two formulas, shown in (1.2), is that both types of forces (gravitational and the electrostatic) are proportional to the inverse of the square of the distance, r, between two masses, m_1 and m_2, or two charges, q_1 and q_2. The constants that appear in both formulas scale the quantities to match empirical phenomena.

$$F_{grav} = G\frac{m_1 m_2}{r^2} \quad \text{and} \quad F_{el} = k\frac{q_1 q_2}{r^2} \tag{1.2}$$

In contrast, the other three types of tool migration, despite prominent examples, are less explored in regard to their general features. One prominent example of tool-transformation is the development of game theory to be used in evolutionary biology.

3. The Migration of Game Theory From Social Sciences to Biology

In this section, I show in what sense the novel use of game theory in evolutionary biology, which is now known as 'evolutionary game theory' ('EGT') can be considered as a tool-transformation. I should mention that my account of the migration of game theory in this paper is not meant to address all the limitations of both game theory and EGT in their respective disciplinary contexts. Instead, the purpose of this account is to show that one *can* detect patterns of migration that have epistemic implications by focusing on the target profile and usage profile of a research tool.

3.1 Game theory in social sciences

Game theory was initially formulated to mathematically model strategic interactions between intelligent, rational agents. In game theory, a game is defined as an interaction between two or

more players in which each player's payoff (e.g., profit) is affected by the decisions made by other players. Typically, such a game assumes both *perfect information* and *common knowledge. Perfect information* assumes that all players know the entire structure of the game (all moves and all payouts) as well as all previous moves made by all players in the game (if it is an iterated or multi-move game). *Common knowledge* is the assumption that all players know that all players have perfect information, and that all players know that all players know that all players have perfect information, and so on. That is, *common* knowledge concerns what players know about what other players know. Moreover, the players also recognize that all players are cognizant that all players are rational, i.e., there is common knowledge of the game and of the *unbounded rationality* of all players. As such, all players will act in the way that takes all other players' potential moves into account in order to maximize their odds of winning. In addition to these assumptions regarding the players of a game, the structure of a game, which refers to the combinations of each move and its payout, is usually summarized in a 'payoff matrix.' Typically, an analysis of a game aims to find out its 'solution,' a unique Nash equilibrium (or sometimes equilibria) of the game.

Game theory has been used in economics, as well as other social sciences, to describe, predict, optimize, or explain a variety of human interactions, such as the economic behaviors of firms, markets, and consumers (e.g., Brandenburger and Nalebuff 1995; Casson 1994) military decisions (Haywood 1954) or international politics (e.g., Snidal 1985).

3.2 Game theory in evolutionary biology

Game theory was later used in evolutionary biology, where a game is understood as phenotypes (or heritable traits) in contest. In 1973, John Maynard Smith and George Price borrowed the formalism of a payoff matrix from game theory to mathematically model the evolution of phenotype frequencies in a population of organisms (see Grüne-Yanoff 2011). Their modeling method assumed that phenotypes are in contest with other phenotypes in a population of organisms. For instance, in a Hawk-Dove game, the contest is embodied by organisms with the phenotype of being aggressive and other organisms that are peaceful. In such a context, the payoff of a move is interpreted as the reproductive success of the phenotype (i.e., the number of copies it will leave to the next generation). Moreover, while the terminology such as 'game,' 'payoffs' and the formalism of a payoff matrix can be seen in the novel use of game theory in biology, the solution to a game in evolutionary biology is decidedly different from the Nash-equilibrium. An evolutionary game theoretic analysis typically looks for an evolutionarily stable strategy (ESS), i.e., a distribution of phenotypes in a population that is stable.

3.3 Epistemic implications of tool transformation

It is clear that the target profile of game theory is no longer the same between its established use

in social sciences and its novel use in biology. First, none of the assumptions of *perfect information*, *common knowledge*, and *unbounded rationality* in what is now known classical game theory (CGT) remain in the novel use of game theory in biology. Second, the moves in EGT are heritable phenotypes exhibited by a group of organisms instead of acts available to players. Third, the payoffs in EGT are the reproductive success of the heritable traits. In this sense, the three assumptions concerning the players were stripped away from the tool - as a result of uprooting game theory from social sciences, and the *heritability* assumption about the moves as well as Darwinian fitness interpretation of the payoff were introduced to the tool - as a result of re-situating it to evolutionary biology.

Note that the change in the target profile forces a limitation to the usage profile of the migrated tool. For instance, nullifying the *unbounded rationality* assumption concerning the players, EGT can no longer be used to optimize a game, i.e., discovering the rationally optimal strategy, which is a common use of game theory in social sciences. For instance, in the prisoner's dilemma, the Nash-equilibrium is for both players to defect. This solution is often interpreted as a prescription for the game; the players are irrational not to defect. However, in a Hawk-Dove game, the ESS obviously has no such normative use. Because the 'moves' of being an aggressive type or a peaceful type are not 'chosen,' the idea of there being normatively better or worse choice of moves is therefore questionable. Moreover, the organisms are not assumed to be rational. Thus, while the players in the prisoner's dilemma could be said to be irrational for choosing to cooperate, this sense of normativity does not carry over to the evolutionary game theoretic analysis of the Hawk-Dove game. One would be mistaken to say that it is ‘irrational’ for the doves to be doves. Thus, the change in the target-profile of game theory, especially the stripping away of the *unbounded rationality* assumption, has resulted in how the migrated tool should or should not be used.[1]

Moreover, applying EGT to study social phenomena (e.g., Axelrod 1984) or cultural evolution (e.g., Skyrms 2010) requires a careful re-defining of the terms (such as fitness) so as to avoid misinterpretation. Using EGT in social sciences, which can be considered as a ‘homecoming’ of the migrated tool, is not uncommon. However, the notion of payoffs in EGT refers to, roughly, the overall biological reproductive success of a group of organisms that exhibit a phenotype. Obviously in a social context, reproductive success of the members of some group is not, very often, the feature of interest. A careful reinterpretation of payoffs is thus needed in every analysis to prevent misleading conclusions.

[1] Of course, a more interesting prescriptive use of the ESS of a Hawk-Dove game might be, for example, to manage ecosystems for optimal predator-prey balance. Nevertheless, it should be noted that a justification for this type of prescriptive use of EGT would require further analysis because it is apparently not be derived from CGT.

To generalize, this example suggests that at least in some cases, a change in the target profile requires a corresponding change in the usage profile, or failure of producing genuine knowledge may follow. So far, I have shown that a solution of an ESS analysis may not be interpreted as an optimization to a Hawk-Dove game. Applying EGT to study social phenomena also requires careful treatment to the notion of payoff. Now if, hypothetically, some researcher were to make either of these two mistakes, his or her novel use of the tool would have been classified as tool-transfer - the novel use changes only the target profile without also changing the usage profile. It suggests that in some cases, tool-transformation may not be as risky as tool-transfer. I will come back to the issue of tool-transfer after some remarks related to the migration of game theory.

4. Contributions of the Tool Migration Analysis

The tool migration typology and its focus on tracking both similarities and differences meets the needs to sharpen discussions concerning inter- or cross-disciplinary use of research tools. Current literature seems to lack a framework to capture important, relational characteristics of the research tools that appear in multiple disciplinary contexts. For instance, 'tool-transformation' captures significant differences in details between CGT and EGT without losing sight of the contextual relationship between the two. In contrast, other terms in the literature, such as 'imports' or 'transfers,' fall short of doing so.

'Imports' signals the importation of research tools from a foreign discipline. In contrast, 'transfers' refers to the use of a scientific model, which was established to study phenomena of one domain, to study phenomena of a different domain. Neither term captures the migration of game theory to biology. As Grüne-Yanoff argues,

> [B]iologists constructed the more sophisticated formal [evolutionary game theoretic] concepts themselves. One could speak of the import of formal concepts only with respect to very basic notions such as strategies or pay-off matrices, and it may be more appropriate to refer to formal inspirations rather than imports or transfers in these contexts. (2011, 392)

Moreover, I have suggested that a change in a tool's target profile without a corresponding change in the tool's usage profile *may* lead to misinterpretation and hence misuse of the tool. If this observation is generalizable, which is debatable, then it follows that cases of tool-transfer are epistemically riskier than cases of tool-transformation. On the other hand, if this observation applies only to some cases, it nevertheless reveals at least two epistemic implications concerning tool migration: 1) when the target profile changes, one must be careful not to draw conclusions that might be natural in the old context but may not make sense within the new context, given the new target, and 2) sometimes a change in target profile can, force a change in usage profile. Potentially failing to recognize when these changes occurred in a migration leads

to risky uses of the migrated tool.

Morgan (2011) has argued that while not all scientific knowledge travels far, those that travel with integrity (i.e., maintaining their content more or less intact during its travels) and travel fruitfully (i.e., finding new users or new functions) are considered to be traveling well. It is relatively easy to quantify the latter feature – one needs to look at just the number of a tool's novel applications. However, determining whether a tool has traveled with integrity is not straightforward. As a starting point, this proposed tool migration framework—especially its distinction between the target profile and the usage profile of a tool—provides a starting point that is crucial for assessing the integrity of a migrated research tool. With this framework, one may discover more patterns of tool migration that impact the epistemic integrity and, consequently, effectiveness of a migrated research tool in a foreign discipline.

5. Conclusion

I have argued that mathematical constructs used in science can be viewed as research tools and their cross-disciplinary novel use as tool migration. I have also argued that making novel use of established tools has its risks, but such an implication is not meant to deter cross-disciplinary sharing of tools. Indeed, certain important breakthroughs in the history of science are due to creative, unconventional, uses of research tools (e.g., the use of Fourier's mathematical treatment of heat to study electrostatics [Thomson 1842] or the use of Faraday's mechanical model of fluid motion to model the electromagnetic field [Maxwell 1861]). Versatile research tools are not rare in science. A framework of tool migration aims to offer not only a useful terminology to characterize the diverse landscape of their versatility but also a groundwork to investigate risky patterns of making novel use of established research tools. Finally, this tool migration approach shows that viewing these constructs as tools whilst acknowledging their representational features (i.e., as captured in their target profile) allows for a balanced understanding of knowledge production - especially those productions that are aided by research tools that have migrated across disciplinary boundaries.

References

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Bailer-Jones, Daniela M. 2009. *Scientific Models in Philosophy of Science*. University of Pittsburgh Press.

Brandenburger, Adam M., and Barry J. Nalebuff. 1995. *The Right Game: Use Game Theory to Shape Strategy*. Harvard Business Review.

Boon, Mieke, and Tarja Knuuttila. 2009. "Models as Epistemic Tools In Engineering Sciences: A Pragmatic Approach." In *Handbook of the Philosophy of Science*, edited by Anthonie Meijers, 687–720. Elsevier B.V.

Casson, Mark. 1994. *The Economics of Business Culture: Game Theory, Transaction Costs, and Economic Performance*. Oxford University Press.

Grüne-Yanoff, Till. 2011. "Models as Products of Interdisciplinary Exchange: Evidence from Evolutionary Game Theory." *Studies in History and Philosophy of Science Part A* 42 (2): 386–97.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Haywood Jr, O. G. 1954. "Military Decision and Game Theory." *Journal of the Operations Research Society of America* 2 (4), 365–85.

Humphreys, Paul. 2002. "Computational Models." *Philosophy of Science* 69 (September): 1–27.

___ 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Knuuttila, Tarja, and Andrea Loettgers. 2014. "Magnets, Spins, and Neurons: The Dissemination of Model Templates across Disciplines." *The Monist* 97 (3). The Oxford University Press: 280–300.

Knuuttila, Tarja, and Andrea Loettgers. 2016. "Model Templates within and between Disciplines: From Magnets to Gases–and Socio-Economic Systems." *European Journal for Philosophy of Science* 6 (3). Springer: 377–400.

Maynard Smith, John, and George Price. 1973. "The Logic of Animal Conflict." *Nature* 246: 15–18.

Maxwell, James Clerk. 1861. "Xxv. on Physical Lines of Force: Part I.–the Theory of Molecular Vortices Applied to Magnetic Phenomena." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21 (139): 161–75.

Morgan, Mary, and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Vol. 52. Cambridge University Press.

Morgan, Mary. 2010. "Travelling Facts." In *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, edited by Peter Howlett and Mary Morgan, 3–39. Cambridge University Press.

___ 2014. "Resituating Knowledge: Generic Strategies and Case Studies." *Philosophy of Science* 81 (5). University of Chicago Press: 1012–24.

Morrison, Margaret. 1999. "Models as Autonomous Agents." In *Models as Mediators: Perspectives on Natural and Social Science*, edited by Mary Morgan and Margaret Morrison, 38–65. Cambridge University Press.

Skyrms, Brian. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press.

Snidal, Duncan. 1985. "The Game Theory of International Politics." *World Politics* 38 (1). Cambridge University Press: 25–57.

Thomson, William. 1842. "On the Uniform Motion of Heat in Homogeneous Solid Bodies and Its Connection with the Mathematical Theory of Electricity." *Cambridge Mathematical Journal* 3 (1842): 71–84.

Van Der Steen, Wim J. 1993. "Towards Disciplinary Disintegration in Biology." *Biology and Philosophy* 8 (3): 259–75.

Representations are Rate-Distortion Sweet Spots

Manolo Martínez (mail@manolomartinez.net)

Abstract

Information is widely perceived as essential to the study of communication and representation; still, theorists working on these topics often take themselves not to be centrally concerned with "Shannon information", as it is often put, but with some other, sometimes called "semantic" or "nonnatural", kind of information. This perception is wrong. Shannon's theory of information is the only one we need.

I intend to make good on this last assertion by canvassing a fully (Shannon) informational answer to the metasemantic question of what makes something a representation, for a certain important family of cases. This answer and the accompanying theory, which represents a significant departure from the broadly Dretskean philosophical mainstream, will show how a number of threads in the literature on naturalistic metasemantics, aimed at describing the purportedly non-informational ingredients in representation, actually belong in the same coherent, purely information-theoretic picture.

1 Information, Shannonian and Dretskean

In what follows I will use a random variable, S, to encode the state the world is in, and another random variable, M, for signals. How should we characterize the information that values of M (i.e., individual signals) carry about values of S (i.e., individual world states)? The most basic quantity with which information theory records dependence among two random variables is the *mutual information* between them. This quantity being an expected value, Dretske (1981, p. 52f) claims, renders it unsuitable for an analysis of representational status, and it should be substituted by notions that record relations between individual states, S_i, and individual signals, M_j. The basic relation which substitutes mutual information in contemporary Dretskean accounts is that of *making a probabilistic difference* (Scarantino 2015): a signal M_j makes a probabilistic difference to the instantiation of a state S_i iff the following *basic inequality* holds:

$$P(S_i|M_j) \neq P(S_i)$$

Nearly all the accounts of information developed in the recent, and not so recent, philosophical literature on this topic are variations on, and attempts to quantify, this inequality. For illustration, in Skyrms (2010, p. 36) the "information in $[M_j]$ in favor of $[S_i]$" is defined as the *pointwise mutual information* (Also *pmi* henceforth) between

state and signal. There is a direct relation between pmis and the basic inequality: the former are nonzero iff the latter is true.

The running thread connecting most prominent contemporary accounts of information is that all there is to Shannon's information theory, at least for the purposes of investigating the nature of representation, is two quantities: the unconditional probability of states and the probability of states conditional on signals, perhaps rearranged as the logarithm of their ratio, or in some other way. Unsurprisingly, from this it is routinely concluded that there is much more to representation than information. This conclusion is premature: informational content in the Dretskean tradition is not by a long shot all there is to information theory. This should not be taken to imply that information is all there is to representation—for one thing, I believe with teleosemanticists (Millikan 1984; Papineau 1987) that teleofunctions have a role to play in a complete theory of representation—but it does mean that no Dretske-style "semanticized information" needs to be recognized, over and above the quantities studied in information theory proper. I will argue that it also means that some prominent proposals as to ways to bridge the information-representation gap are, in fact, unwittingly appealing to informational structure.

In the following section I review two such proposals. My aim is not to argue against them—they are built upon largely correct insights. I will instead aim at showing that a better informed understanding of information provides a way to incorporating these insights in a unified, purely information-theoretic picture.

2 Bridging Information and Representation

2.1 Many-to-One-to-Many Architectures

The first proposal is that it is not enough that representations carry information; on top of that, they must sit in the right place in a certain cognitive architecture. Sterelny (2003), for example, has argued that the emergence of representations is enabled by two prior evolutionary transitions: from "detection" to "robust tracking", on the one hand; from "narrow-banded" to "broad-banded" behavioral responses, on the other. Robust tracking is in essence a *many-to-one* relation between world state and signal: many sensory inputs give rise to one and the same representation. Other theorists have advocated similar architectural constraints on representational vehicles. Famously, Burge (2010) places a great deal of weight on *perceptual constancies* in his characterization of perceptual representation (Burge 2010, p. 413.) This is a variation on Sterelny's idea and, as such, a many-to-one architectural constraint on representational status.

As for broad-banded responses, in these systems a single representation will be flexibly dealt with, resulting in different courses of action, depending on the context where the representation is tokened. Response breadth is in essence a *one-to-many* relation between representational vehicle and output: one representation, many agential outputs.

2.2 Reference Magnetism

A second proposal has been to focus on the entities that should figure in the content of simple representations. The suggestion, typically, is that represented entities should be appropriately *natural*, or *real*. For example, Dan Ryder (2004, 2006) has argued that neurons become attuned to *sources of correlation*. These entities are closely related to Richard Boyd's *homeostatic property clusters* (also HPC henceforth, Boyd 1989): HPC theory identifies natural kinds with clusters of properties which tend to be instantiated together, and such that this frequent co-instantiation is not just a statistical fluke. What Ryder calls sources of correlation are the grounds for these HPC-related frequent co-instantiations—whatever it is that makes them *not* statistical flukes. Ryder claims that many of the representations the brain trades in target sources of correlation. Martínez (2013) and Artiga (forthcoming) have made more general cases that simple representations preferably target HPCs (Martínez), or properties that best explain the co-occurrence of other properties (Artiga).

A similar idea has been explored in an entirely independent line of enquiry starting with Lewis (1983): "among the countless things and classes there are … [o]nly an elite minority are carved at the joints, so that their boundaries are established by objetive sameness and difference in nature. Only these elite things and classes are eligible to serve as referents" (Lewis 1984, p. 227). This is what Sider (2014, p. 33) calls *reference magnetism*.

As I show in section 4, these two ideas, although apparently disparate, are in fact closely related, and the explanatory payback they bring to representation-involving talk depends on their informational underpinnings.

3 Information Theory is a Source-Channel Theory

Philosophy has understood information theory as a mostly *definitional* effort: for all philosophers have typically cared, the theory begins and ends with a presentation of what it takes for one random variable (or the worldly feature it models) to carry information about another. But information theory goes well beyond that. It is, well, a *theory*, and as such it is chiefly composed of claims that are advanced in the hope that they be true about the world.

In a nutshell, the most celebrated results in information theory have to do with specifying how faithful the transmission of information from a source can be, when it happens over a (typically noisy, typically narrow) channel. These results have played absolutely no role in informational accounts of representation.[1] Take, for starters, the idealized depiction of an information-processing pipeline in fig. 1 (*cf.* Cover & Thomas 2006, fig. 7.1)

[1] Two recent philosophical treatments of information that try to redress this neglect are Mann (2018) and Rathkopf (2017).

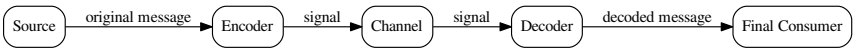


Figure 1: An information-processing pipeline

Here an *encoder* produces a signal as a response to information incoming from a source. This signal goes through a channel and is subsequently decoded, producing a message that is then utilized for whatever purposes downstream. The first thing to note is that the broadly Dretskean ideas about the content of a signal introduced in section 1 only have use for the first two links in this information-processing chain: how signals carry information about a certain original message produced by a source, as depicted in fig. 2. In fact, in information theory the main action happens immediately after that: a source is producing stuff, and we want that stuff to *go through a channel*. Information theory is mainly about providing theoretical guarantees of faithfulness in transmission, given the rate of the channel. We can think of this rate as the number of bits it provides for the encoder to use in the signal. If, say, the rate is 2 bits per use of the channel, this means the encoder can use up to 2 bits to construct the signal and be sure that it can pass unscathed through the channel and on to the decoder.

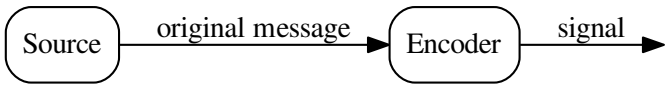


Figure 2: The information-processing pipeline in the Dretskean tradition

In typical cases of representation, channel rate is consistently smaller than ideal. Consider animal alarm calls. Vervet monkeys, for example, are typically described as being able to produce three different, discrete kinds of calls (Seyfarth, Cheney & Marler 1980a, 1980b) that are usually taken to be associated with the presence of leopards, eagles and snakes respectively. Obviously, the entropy of the relevant aspects of the environment that prompt the production of a call (think of all the possible patterns of approach of these predators, for example) vastly outstrip the rate of a channel, which consists in the production of just one out of three possible signals. This means that loss in communication is inevitable. Alarm calls, and for analogous reasons representations in general, are all about *lossy transmission*.

The way in which information theory deals with lossy transmission is by defining a *distortion measure* (Cover & Thomas 2006, p. 304) that gives a score to a pair composed of a certain original message M, and the decoded version thereof, $\hat{M}$. In what follows I

will be using the *Hamming distortion* which simply adds 1 to the distortion when the bits in the original and decoded signals (which we can assume to be binary strings) do not coincide, and 0 otherwise, then normalizes. So, for example, the Hamming distortion between an original signal $M = 010011$ and a decoded signal $\hat{M} = 100010$ is ⅜, because the first, second, and last (a total of 3) bits have been decoded incorrectly, and there are 6 bits in total.

The central result in this so-called *rate-distortion theory* approach to lossy transmission is that there is a *rate-distortion function*, $R(D)$, which gives the minimum rate at which any given distortion is achievable. The actual mathematical expression of the rate-distortion function need not detain us here (see Cover & Thomas 2006, p. 307, theorem 10.2.1), but it is such that the *Blahut-Arimoto* algorithm (Blahut 1972; Arimoto 1972) allows us to calculate it easily.

The main thesis of this paper is that representations belong in information-processing pipelines whose rate-distortion function has *sweet spots*: by this I mean points in the rate-distortion curve such that the usefulness of increasing the rate of the channel past those points is much smaller than before reaching them. Moreover, the encoding-decoding strategies that make use of these representations tend to live in the vicinity of those sweet spots. I submit that it is these information-theoretic properties that the conditions on representation discussed in section 2 try to get at.

To see how rate-distortion analyses work let's start by looking into a source that models a series of fair-coin tosses: this random variable would have two values, *heads* and *tails*, with associated probabilities $P_{heads} = P_{tails} = .5$). Using the Hamming distortion as our target distortion measure, if the coin lands heads (tails) and the decoded message is tails (heads) the distortion is 1, otherwise 0. The Blahut-Arimoto algorithm allows us to draw the rate-disortion curve, in fig. 3. Here the blue line is the rate-distortion curve. It intersects the x-axis at 1.0 bits (the entropy of the source) and it intersects the y-axis at 0.5 (the lowest average distortion one can achieve when the channel is closed.) The red line gives a measure of how steep the blue line is at any given point—in particular, the absolute value of the slope of the blue line. The higher the red line, the steeper the blue line.

The situation this setup is modeling is one in which a single cue is present or absent, and a signal tries to keep track of whether it does. This is precisely the kind of situation where many theorists (certainly Sterelny and Burge, for the reasons reviewed in 2.1) would see the postulation of representations as entirely idle—see, e.g., Schulte's vasopressin example in his Schulte (2015). In agreement with the idea that postulating representations here is idle, there is not much structure to the rate-distortion curve corresponding to this setup: reading the chart from right to left, increasing the rate makes the achievable expected distortion go smoothly down, until the rate hits the entropy of the source, at which point the achievable distortion is zero.That's about it.

Let's now model one kind of situation in which there is a reasonably wide consensus that representations make an explanatory contribution: vervet-monkey alarm calls, as reviewed above. In the model, the source—the situation the information-processing pipeline is dealing with—randomly makes members of two natural kinds (we can think

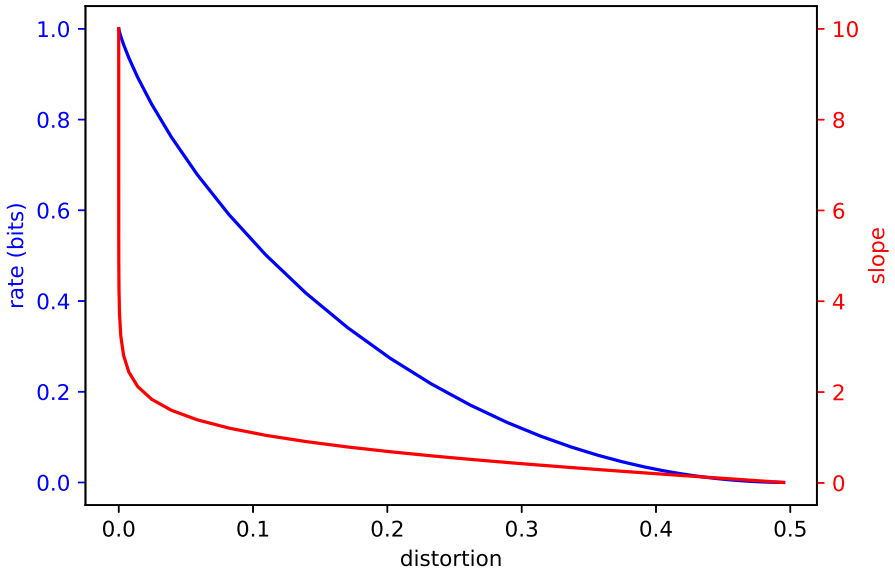


Figure 3: The rate-distortion function for a coin toss

of them as two different predators) be or not present at any given time, independently from one another. This intends to mimic the situation vervet monkeys face, where snakes, leopards and eagles show up or not, more or less at random.

These natural kinds are modeled as homeostatic property clusters (see section 2.2 above). In order to derive a explicit probability distribution for the source out of this qualitative description, the two HPCs are in their turn represented by two Bayesian networks, each with a parent node and four children (see fig. 4.) Each of the nodes stands for a property; if the node is *on* it means the corresponding property is instantiated; if it is *off* it means it is not. In the model, children nodes replicate noisily the state of their parent. Thus, e.g., if the parent is *on* (if the corresponding property is instantiated) each child property will have a .95 chance of being instantiated too; if the parent is *off* the probability for each children of being instantiated is .05. The unconditional probability of instantiation for the two parent nodes is .5.

In the model, the source produces a binary string, with each member of the string being 1 if the corresponding node is on, and 0 if it's off. This signal is encoded, goes through a channel, and is then decoded at the other side. The target distortion measure is the Hamming distortion. Fig. 5 plots the rate-distortion curve for this model.

This curve is very different from the one in fig. 3: there is a clear "sweet spot"—a sudden drop in the usefulness of extra rate, see the red curve—when the system hits a rate of 2 bit/use. I.e, there is, in a certain principled sense, an optimal level of lossy compression; a way to set up an encoding-decoding strategy that recover most of what's going on in

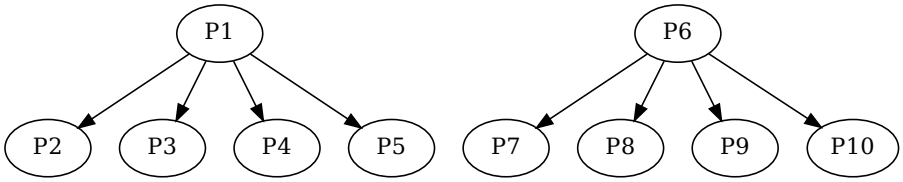


Figure 4: Two natural kinds

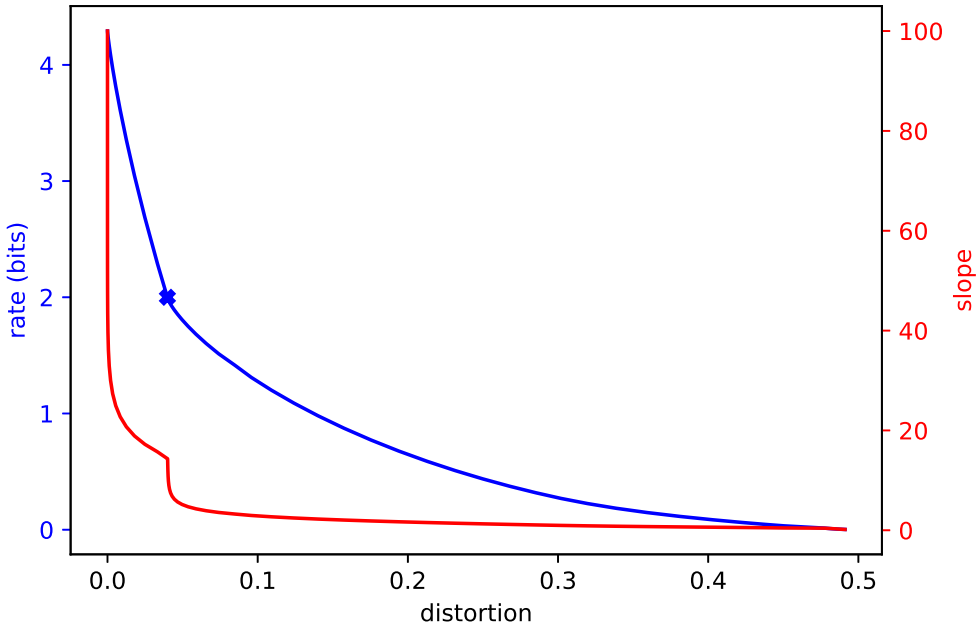


Figure 5: A sweet spot in the rate-distortion function

the world of relevance to the information-processing system, even through a very severe, 2 bit bottleneck. I claim that this is no coincidence. Our representation-attributing practices gravitate towards this kind of situations.

To see how sweet spots in rate-distortion curves and representations are related, consider now what an optimal encoding-decoding strategy would look like. That is, how should the encoder encode the information coming from the source, and how should the decoder decode the signal coming from the encoder, so that the resulting expected distortion between original and decoded signal is the minimum achievable, at the sweet spot?

Optimal Encoding Strategy: First divide the incoming signal in two halves, one corresponding to properties P_1 through P_5; the other corresponding to properties P_6 through P_{10}.

If there is a majority of 1s in the first half of the original signal set the first bit of the signal to 1. Otherwise set it to 0. Ditto for the second half of the original signal and the second bit of the signal.

Optimal Decoding Strategy: If the first bit in the incoming signal is 1, set the first half of the decoded signal to 11111. Otherwise, set it to 00000. Ditto for the second bit and the second half of the decoded signal.

How should we interpret what encoder and decoder are doing here? A natural way is this: they are using the presence or absence of properties in an HPC cluster as diagnostic of the presence or absence of the underlying natural kind—this would be the encoding part—and then taking the resulting signals as representing the presence of a paradigmatic instance of the kind, one that has all the properties in the cluster—this would be the decoding part. HPC kinds being what they are, frequently the first half of the incoming signal will resemble the paradigmatic presence of the first kind (11111) or its paradigmatic absence (00000), and the same will happen with the second half and the second kind. That is why this encoding-decoding strategy works so well.

In describing this optimal strategy I have helped myself to representational vocabulary; it has been useful in order to explain how the strategy works, and how come that behaving in this particular way achieves low distortion at low rates: it is because each of the two bits in the signal is caused by, and causes, behavior that is optimally attuned to the probabilistic structure of each of the two natural kinds in the model world, respectively. Nothing going on in this system falls outside the purview of Shannonian information theory—of information theory *tout court*, so at least in this kind of cases representational talk depends on no non-informational fact.

We can now understand better what's lacking in the philosopher of mind's information-theoretic tookit: it is entirely possible, and computationally trivial, to calculate, e.g., Skyrms's pmi between each of the possible signals (00, 01, 10 and 11) and each of the possible world states (all 1024 of them, from 0000000000 to 1111111111). Doing so would leave us with 4 vectors (one for each signal) with 1024 entries each (one for each world state.) First, this is an unwieldy collection of numbers, which doesn't bring out the relevant structure. For example, if the probability of children nodes being *on* conditional on their parent being *on* was .96 instead of .95 the rate-distortion curve

would be qualitatively identical, with a sweet spot in exactly the same place, yet most numbers in the Skyrmsian informational content vectors would change. Second, and most important, nothing in those 4096 numbers allows us to infer the presence of a sweet spot. The relevant information is simply not there, depending as it does on a distortion measure which is not used in computing Skyrmsian informational contents.

If this is approximately right, the question about what makes representational talk explanatory is readily answered: saying that a certain vehicle is a representation conveys something quite specific about its informational context. It says that the vehicle is part of an encoding-decoding strategy that exploits a sweet spot in a rate-distortion curve—where the curve is in turn fixed by the probabilistic structure of the world, and the target distortion measure. This, in less technical terms, translates to saying that the vehicle is summarizing *relevant* (this is where the distortion measure comes in) aspects of the current situation in an optimal, if lossy, manner, made possible by *how the world* is (this is where the probabilistic structure of the world comes in.) This explication of the explanatory contribution of representations can be turned into an explicit answer to what makes something a representation—an answer, that is, to what Artiga (2016) calls the metasemantic question.

The Rate-Distortion Approach: A signal, S, in a certain information-processing pipeline, P, is a representation if the following two conditions are met:

Existence: There are sweet spots in the rate-distortion curve associated with P.

Optimality: S is produced as part of an encoder-decoder strategy that occupies the vicinity of one of these sweet spots.

So, *pace* Dretske, the core information-theoretic notions of entropy, rate, distortion, etc. *can* provide invaluable insight into the representational status of individual signals. If the rate-distortion approach is on the right track, those information-theoretic notions, through the existence condition, specify the kind of setup where representations live, which then the optimality condition can use to provide a criterion for the representational status of individual signals.

I offer the foregoing discussion as a preliminary case for the rate-distortion approach to representation: it shows how postulating representations is explanatory, even if these representations depend just on (Shannon) information. It illuminates the difference in representational status between cue-driven examples, such as Schulte's vasopressin; and vervet alarm calls, and other similar examples. To complete my case I now show how the ways to bridge the gap between natural and nonnatural information discussed in section 2 can be seen as unwitting attempts to get at rate-distortion sweet spots.

4 There is no Gap to Bridge

What does it take for the existence condition to be met? That is to say, what circumstances result in sudden drops in the slope of the rate-distortion curve? We have seen one such family of circumstances: if the pattern in which properties are instantiated

in the source is noisily replicated in a cluster then sudden drops are to be expected: distortion will decrease with rate up to the point where all the main sources of variation in property instantiations are accounted for, and all that remains is the residual noise in instantiations within each cluster. Take a look again at figs. 4 and 5: to describe this source we basically need enough rate to account for the two main sources of variation: P_1 and P_6. This is not all there is to the world, because it's possible for the other properties to (fail to) token independently of their parent, but the unlikeliness of these departures makes the extra rate comparatively less useful.

Noisy replication of property instantiations is at the core of the HPC theory of natural kinds, as we saw above. This means that, in general, the presence of HPC natural kinds in a source will create sweet spots. This opens a line of argument in favor of reference magnetism from information-theoretic premises: reference magneticsm should be seen as making a point about the kind of probabilistic structure that an information-processing pipeline must be attuned to, if signals are to effect the kind of optimal lossy compression that underlies our representation-attributing practices. Reference magnetism is just a way of meeting part of the existence condition.

Regarding the suggestion, by Sterelny, Burge and others, that representations inhere preferably on signals sitting in a one-to-many-to-one pipeline, I submit that the many-to-one aspect of this suggestion aims at meeting the optimality condition; the one-to-many aspect, together with reference magnetism, aims at meeting the existence condition.

The first thing to note here is that the *Optimal Encoding Strategy* presented above enforces what Sterelny calls robust tracking and Burge calls constancy: the strategy consists in considering all properties coming from each of the two clusters and setting the relevant bit to 1 only if a majority of those properties are instantiated. That is, the encoder is taking a multiplicity of configurations (e.g., the first half of the incoming signal being 00111, 01011, 10111, etc.) to a single output: the first bit of the signal being 1. Furthermore, that part of the signal will be decoded as 11111: from there on, the system downstream will treat whatever is out there in the world as a paradigmatic member of the first kind. The system is recovering the presence of a natural kind out of many different, noisy instantiation patterns. This is a clear instance of constancy. Suppose that the encoder, insted of being many-to-one, depended on a single cue; say, suppose it set the first bit to 1 if one of the children properties (say, P_2) was instantiated, and to 0 otherwise. In such a cue-driven setup, the best encoder-decoder arrangement possible is marked by the blue circle in fig. 6. This has double the distortion than than the optimal encoding (marked by the blue cross) which sits right on top of the optimal rate-distortion curve. This cue-driven system would not meet the optimality condition, which means that a many-to-one architecture is instrumental to meeting it.

Finally, the target distortion measure in the information-processing pipeline can be seen as that which Sterelny's one-to-many condition on representation is actually tracking. Using, for example, the Hamming distance as a distortion measure is tantamount to assuming that all of the properties of the natural kinds are relevant for downstream processing. One natural way in which this may happen is when the agent is to respond flexibly to the presence of the natural kind: in different contexts or states different properties of the kind might be relevant and, for example, the presence of a tree might

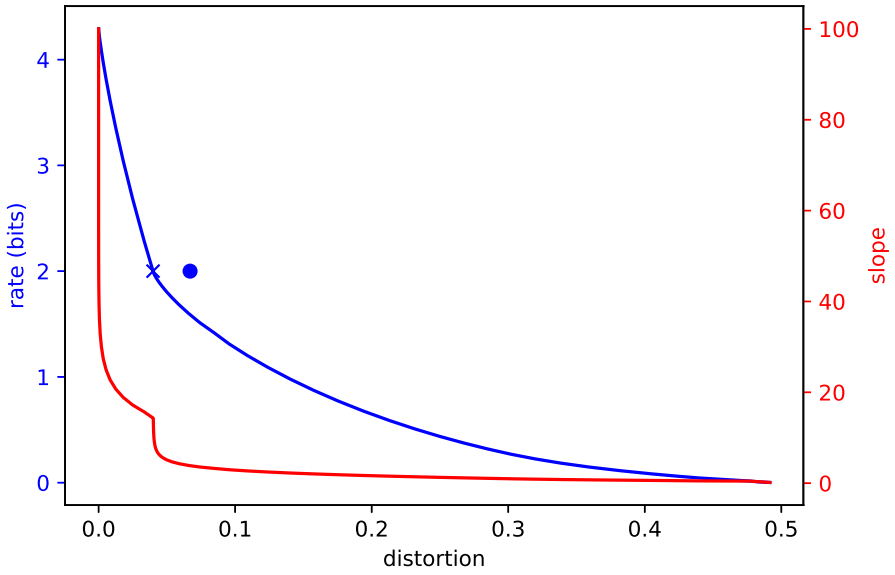


Figure 6: Cue-driven encoding

be sometimes relevant to behavior because it bears fruit (if the agent is hungry) and some other times because it has a dense cover (if the agent is looking for shelter.)

Caring about all (or many) properties of the kind is what makes the rate-distortion curve display a sweet spot. If, instead, the agent has a rigid, stereotyped response to the presence of members of the kinds—that is, if it only cares about the presence of one property, which is the property that makes that rigid behavioral response fitness-conducive, then the curve is as presented in fig. 7. Rigid behavioral responses make the probabilistic structure of the kinds largely irrelevant. As a result, the system behaves as if a coin were tossed, where heads would mean that the target property is tokened, and tails that it is not. This arrangement does not meet the existence condition. Sterelny's broad-banded responses are, again, a way of getting at rate-distortion sweet spots.

References

Arimoto, S 1972, 'An algorithm for computing the capacity of arbitrary discrete memoryless channels', *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20.

Artiga, M forthcoming, 'Beyond Black Spots and Nutritious Things: A Solution to the Indeterminacy Problem', *Dialectica*.

Artiga, M 2016, 'Liberal Representationalism: A Deflationist Defense', *dialectica*, vol.

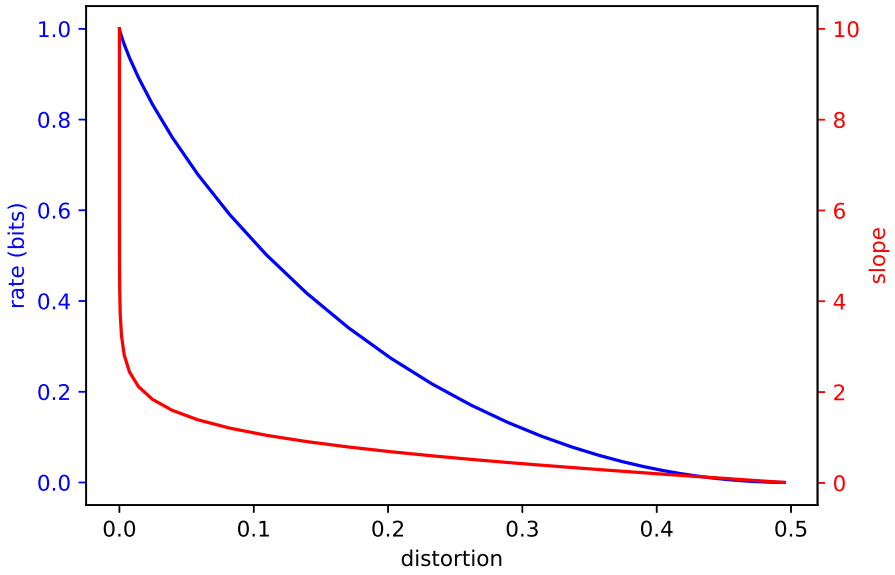


Figure 7: Rigid behavioral response

70, no. 3, pp. 407–430.

Blahut, R 1972, 'Computation of channel capacity and rate-distortion functions', *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473.

Boyd, R 1989, 'What Realism Implies and What It Does Not', *Dialectica*, vol. 43, no. 1-2, pp. 5–29.

Burge, T 2010, *Origins of objectivity*, Oxford University Press.

Cover, TM & Thomas, JA 2006, *Elements of Information Theory*, New York: Wiley.

Dretske, F 1981, *Knowledge and the Flow of Information*, The MIT Press.

Lewis, D 1983, 'New work for a theory of universals', *Australasian journal of Philosophy*, vol. 61, no. 4, pp. 343–377.

Lewis, D 1984, 'Putnam's paradox', *Australasian Journal of Philosophy*, vol. 62, no. 3, pp. 221–236.

Mann, SF 2018, 'Consequences of a Functional Account of Information', *Review of Philosophy and Psychology*, pp. 1–19.

Martínez, M 2013, 'Teleosemantics and Indeterminacy', *Dialectica*, vol. 67, no. 4, pp.

427–453.

Millikan, R 1984, *Language, Thought and Other Biological Categories*, The MIT Press.

Papineau, D 1987, *Reality and Representation*, Basil Blackwell.

Rathkopf, C 2017, 'Neural information and the problem of objectivity', *Biology & Philosophy*, vol. 32, no. 3, pp. 321–336.

Ryder, D 2006, 'On Thinking of Kinds', in G Macdonald & D Papineau (eds), *Teleosemantics*, Oxford University Press, pp. 1–22.

Ryder, D 2004, 'SINBAD Neurosemantics: A Theory of Mental Representation', *Mind & Language*, vol. 19, no. 2, pp. 211–240.

Scarantino, A 2015, 'Information as a probabilistic difference maker', *Australasian Journal of Philosophy*, vol. 93, no. 3, pp. 419–443.

Schulte, P 2015, 'Perceptual representations: A teleosemantic answer to the breadth-of-application problem', *Biology & Philosophy*, vol. 30, no. 1, pp. 119–136.

Seyfarth, RM, Cheney, DL & Marler, P 1980a, 'Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication', *Science*, vol. 210, no. 4471, pp. 801–803.

Seyfarth, RM, Cheney, DL & Marler, P 1980b, 'Vervet monkey alarm calls: Semantic communication in a free-ranging primate', *Animal Behaviour*, vol. 28, no. 4, pp. 1070–1094.

Sider, T 2014, *Writing the Book of the World*, Reprint edition., Oxford University Press, Oxford.

Skyrms, B 2010, *Signals: Evolution, Learning & Information*, New York: Oxford University Press.

Sterelny, K 2003, *Thought In A Hostile World: The Evolution of Human Cognition*, John Wiley & Sons, Malden, MA.

The Proportionality of Common Sense Causal Claims

Jennifer McDonald

This paper defends strong proportionality against what I take to be its principal objection – that proportionality fails to preserve common sense causal intuitions – by articulating independently plausible constraints on representing causal situations. I first assume the interventionist formulation of proportionality, following Woodward.[1] This views proportionality as a relational constraint on variable selection in causal modeling that requires that changes in the cause variable line up with those in the effect variable. I then argue that the principal objection derives from a failure to recognize two constraints on variable selection presupposed by interventionism: *exhaustivity* and *exclusivity*.

[1] Woodward 2003

1. Introduction

Yablo's principle of proportionality holds, roughly, that something counts as a cause of some effect just in case it includes the appropriate degree of causal information.[2] Proportionality has been put to various philosophical uses, such as a proposed solution for the causal exclusion argument, and as a justification and explanation of the dependence on high-level causal explanations in the special sciences. However, the precise formulation of such a principle has proven to be controversial.

I take the most promising formulation to be an interventionist one, following Woodward.[3] Such a formulation defines proportionality as a relational constraint on variable selection in causal modeling. In this paper, I argue that this formulation works well as it is – contra Franklin-Hall (see 2016) – so long as we recognize two independently plausible background requirements on variable selection. I call these *exhaustivity* and *exclusivity*. Exhaustivity holds that a variable must take at least one of its values. Exclusivity holds that a variable can take at most one of its values. Both constraints are relative to, and thereby help to make explicit, the modal assumptions implicit in causal inquiry.

Finally, with these requirements in place, I defend proportionality against its principal objection: that it fails to preserve fundamental causal intuitions. I demonstrate how this concern derives from a failure to recognize and integrate the modal assumptions implicit in causal inquiry, in tandem with an inappropriate use of variables to represent causal situations.

2. Interventionism

The formulation of proportionality that I endorse comes directly from Woodward, and is defined in terms of his interventionist account of causation. Interventionism expands on the intuition that causal claims provide

[2] Yablo 1992
[3] Woodward 2003, 2008a, 2008b, 2010, 2016

manipulability information. If *X* causes *Y*, then manipulating or changing *X* is a way of manipulating *Y*. It then exploits the language of causal models to identify and articulate different causal relations of interest. A causal model can take a variety of forms, such as graphical, potential-outcome, and structural-equations models.[4] However, I'll restrict discussion of causal models in this paper to graphical models. A graphical model is, essentially, a set of variables – representing the causal relata – and a directed binary relation between them – representing causal influence.

Interventionism then defines the notion of an *intervention* on a system. An intervention, *I*, first must directly change the value of some variable, *X*, in such a way that it breaks the dependence that *X* may have had on other variables in the system. Second, *I* must be designed in such a way that any change in the effect variable, *Y*, will be the direct result of *X* and not of *I* itself. Finally, *I* must be wholly independent of other possible causes of *Y*, whether such causes are represented by the given model or not. A more precise formulation than this won't matter for the purposes of this paper.[5]

With this in place, the interventionist then defines a basic notion of cause, which corresponds most closely with the intuitive notion of *causal relevance*:

> (Principle M) *X* causes *Y* iff there are background circumstances B such that if some (single) intervention that changes the value of *X* (and no other variable) were to occur in B, then Y would change. (Woodward 2003, 222)

That is, in order for *X* to be a cause of *Y*, the change in *X* from one value to another as the result of an intervention corresponds to the change in *Y* from one value to another, given some fixed set of background parameters. Various kinds of causal relations are then captured by refinements on this basic notion. Due to

[4] See Greenland and Brumback 2002 and Hitchcock 2009 for overviews of causal models.
[5] See Woodward 2003, chapter 3, especially 98

the irrelevance of these and further details to my argument, I'll leave my overview of interventionism here.[6]

3. Proportionality as Relational Constraint on Variable Selection

Interventionism places variables front and center in how we represent and inquire into causation. Thus, more needs to be said about the criteria for variable selection. Although the variables can be taken to represent different things, I will assume throughout that the set of values of a particular variable represents a set of properties – constrained by a given property type – that are possibly instantiated by some particular thing. The assumed causal relata of this paper will therefore be property instantiations.

This paper addresses two questions relevant to variable selection: (i) What determines the range of values that a variable can take? (ii) At what level of description should the values of the variables be? Proportionality has been proposed as an answer to (ii). However, after laying out the proposal, I'll go on to argue that while (ii) can be answered by the principle of proportionality, it can only do so alongside an appropriate answer to (i). One aspect of such an answer is that the background modal context determines the range of values that a variable takes.

Constraints on variable selection can be divided into two kinds: relational constraints and non-relational constraints. *Relational constraints* pertain to the extrinsic nature of the variables in a causal model, to how "variables relate to one another." (Woodward 2016, 1056) One example of such a constraint is stability.[7] *Stability* is the persistence of the causal relation between a cause variable and an effect variable, despite changes in the background conditions. The more changes such a relation can survive, the more stable it is.

[6] See Woodward 2003, chapter 2, especially section 3
[7] See Woodward 2010, 2016

Proportionality is just such a relational constraint. It holds that changes in a cause variable should line up with changes in an effect variable. Intuitively,

> Proportionality has to do with whether changes in the state of the cause 'line up' in the right way with changes in the state of the effect and with whether the cause and effect are characterized in a way that contains irrelevant detail. (Woodward 2010, 287)

Take Yablo's pigeon example.[8] Sophie the pigeon is trained to peck at red things and only at red things. She then pecks at a paint chip, which is a particular shade of red – scarlet. Which of the following is causally relevant to Sophie's pecking: the chip's being red or the chip's being scarlet?

When translated into interventionist terms, this becomes a false dichotomy. Take the variable, *P*, to be a variable representing whether the pigeon pecks or not. It can take the values: {*peck*, *not-peck*}. Now consider two alternative variables for representing the property-instantiations of the paint chip: the variable, *R*, which can take the values {*red*, *not-red*}, and the variable, *T*, which can take the values {*taupe, scarlet, cyan, mauve, crimson,* etc.}, where 'etc.' stands for all other physically possible colors at the same grain as those already made explicit. According to Principle M, the causal model in which *R* stands as causally relevant to *P* is just as accurate as one in which *T* so stands. In the *R* model, *R* is causally relevant to *P* because an intervention on *R* that changes its value from *not-red* to *red* changes *P*'s value from *not-peck* to *peck.* In the *T* model, *T* is causally relevant to *P* because an intervention on *T* that changes its value from *taupe* to *scarlet* changes *P*'s value from *not-peck* to *peck*.

Interventionism therefore doesn't ask the question, which variable stands in a causal relation to *P*? For, the answer is 'both'. *R* and *T* are each causally relevant to *P*. But, this doesn't mean that their respective relationship to *P* is the same. *R* is *proportional* to *P*, while *T* is not. All of the changes in *R* line up with changes in *P* – every intervention on *R* corresponds to a change in *P*. But only some of the

[8] Yablo 1992

changes in *T* line up with those in *P* – only certain interventions on *T* correspond to changes in *P*. The intervention that changes the value of *T* from *taupe* to *cyan*, for example, will not change the value of *P*.

Woodward defines proportionality more explicitly as,

> (P) There is a pattern of systematic counterfactual dependence (with the dependence understood along interventionist lines) between different possible states of the cause and the different possible states of the effect, where this pattern of dependence at least approximates to the following ideal: [it] should be such that (a) it explicitly or implicitly conveys accurate information about the conditions under which alternative states of the effect will be realized and (b) it conveys only such information – that is, the cause is not characterized in such a way that alternative states of it fail to be associated with changes in the effect. (2010, 298)

There are two views on what this difference between variables like *R* and *T* means. The first takes proportional variables such as *R* to represent genuine causes, while non-proportional variables such as *T* represent merely causally relevant factors. Proportionality is thereby considered a necessary constraint on causation. Call this *strong proportionality*.[9] The second view takes proportionality to be a merely pragmatic constraint on causal explanation.[10] Call this *weak proportionality*. Throughout this paper, I assume and defend strong proportionality.

4. Non-Relational Constraints: Exhaustivity and Exclusivity

Non-relational constraints, on the other hand, pertain to the intrinsic nature of the variables in a causal model. These constraints "can be applied to variables, individually, independently of how they relate to other variables." (Woodward

[9] See List and Menzies 2009; Menzies and List 2010; and Papineau 2013
[10] See Woodward 2015; Shapiro and Sober 2012; McDonnell 2017; and Weslake 2013, 2017

2016, 1057) One example is *metaphysical naturalness*, which requires that variables pick out only natural properties, on some understanding of 'natural'.[11]

What I propose to call the exhaustivity and the exclusivity constraint are similarly non-relational constraints. Take exhaustivity first. The *exhaustivity constraint* requires that a variable's values capture the entire range of relevant possibilities for whatever type of thing the variable represents. An exhaustive variable is one that must take one of its values, given whatever background modal constraints are in place.

Since I've restricted this discussion to variables whose values represent the property instantiation of some target object, I can define exhaustivity in more precise terms. *Exhaustivity* is the constraint on a variable in a causal model that holds that its values must jointly represent the range of possibilities of property instantiation by the given object for the given property-type. If the property-type is a color, for example, then the values must somehow exhaust the color spectrum. This can be done quite simply with a binary variable that can take the values: {*some particular color*, *not-(that particular color)*}.

Next, the *exclusivity constraint* holds that the values of a given variable should be such that any one excludes all the others. Woodward references exclusivity when he writes,

> When considering the values of a single variable, we want those values to be logically exclusive, in the sense that variable *X*'s taking value *v* excludes *X*'s also taking value *v'*, where $v \neq v'$. (2016, 1064)

In other words, if two things are not exclusive – if they could occur together – then they should be represented by distinct variables. While exhaustivity holds that a variable should take *at least* one of its values, exclusivity holds that a variable should take *at most* one of its values.

[11] See Lewis 1983; Menzies 1996; Paul 2000; and Franklin-Hall 2016

8

Importantly, exhaustivity and exclusivity are each relative to a background modal context. In possible worlds terminology, the modal context is the set of possible worlds relevant to the truth of the counterfactual that captures the causal claim. It can be described as a set of worlds, or perhaps more succinctly as a list of background assumptions that define such a set. These assumptions can include any constraint that operates in a law-like fashion.

For example, the causal claim, "The chip's being scarlet caused the pigeon to peck," corresponds to the counterfactual, "Had the chip not been scarlet, the pigeon wouldn't have pecked." The modal context of this claim and corresponding counterfactual is the set of possible worlds that determines whether the counterfactual is true. So, if this claim and counterfactual are meant to represent a *specific* causal situation near a local paint chip factory that specializes in just the colors scarlet and cyan, and no others, then the relevant set of possible worlds will be constrained to those in which the paint chip takes one of the two factory colors – cyan or scarlet. In this context, the variable, *C,* that can take the values {*cyan, scarlet*}, is an exhaustive variable. Further, given this set of worlds, the counterfactual is true.

If instead these are meant to represent any *general* causal situation involving paint chips and a red-pecking pigeon, then the relevant set of possible worlds will be more inclusive, including all worlds in which the paint chip takes any color within the color spectrum. *C* is not exhaustive relative to this more inclusive modal context. But the variable *T*, from before, is. Given this more inclusive set of worlds, the counterfactual is false, since the pigeon will peck in response to shades of red other than scarlet.

A point of note here is that the constraints of exhaustivity and exclusivity are indeed non-relational constraints in the sense previously defined. Although they are relative to the modal context, they are *not* relative to other variables in the model. They are properties of a variable taken independently as a representation of the target scenario.

I hold that causal models successfully represent causal situations in part by requiring exhaustive and exclusive variables. Proportionality, defined in terms of causal models, also requires exhaustive and exclusive variables. A significant upshot of this is that the proportional cause is not only relative to the target effect variable, but also to the background modal context.

5. Interventionist Proportionality Does the Trick

Franklin-Hall contends that Woodward's formulation of proportionality doesn't successfully prioritize intuitively proportional causal relata, such as red in the pigeon example. However, as I'll argue, presupposing my notion of exhaustivity corrects for this objection.

Franklin-Hall argues that proportionality as laid out in section 3 is inadequate for capturing the kind of causal explanation we're looking for. To do so, she calls upon Sophie and her paint chip. She then introduces a comparison between the causal variable, *R*, that can take the values: {*red*, *not-red*}, (as above), and a variable, *C*, that can instead take the values: {*cyan, scarlet*} (as above). *R*, as before, is proportional to, and therefore a genuine cause of, *Y*. But, she argues, *C*, too, is proportional to *Y*, since every possible intervention on *C* changes the value of *Y*. An intervention on *C* that changes its value from *cyan* to *scarlet* changes *Y* from *not-peck* to *peck*, and an intervention that changes *C*'s value from *scarlet* to *cyan* changes *Y*'s value from *peck* to *not-peck*. Thus, the changes in *C* line up with the changes in *Y* just as well as the changes in *R* do. The problem, then, is that proportionality, as formulated, is insufficient to its intended task. It fails to privilege a variable like *R* over one like *C*, and so fails to prioritize a causal model that uses *R* over one that uses *C*.

In response to this problem, a natural move would be to find a way to disqualify variables like *C* from the arena. Intuitively, *C* is not the right kind of variable. But, why not? I propose that our aversion to variables like *C* is due to their failure to exhaustively represent the implicit modal context of the situation. The background possibilities relative to the paint chip include the full color spectrum.

Unless the possible color of the paint chip is restricted in some way – by the local factory, for example – then the target object can fail to take one of *C*'s two values. There are other physically possible colors that the paint chip could have – such as beige or olive green – and *C*'s values fail to represent these possibilities.

Relative to the implicit modal context, then, *C* is not an exhaustive variable. The variable, *R*, on the other hand, is exhaustive, since the object must take one of *R*'s two values. By requiring exhaustive variables, *C* is discounted as a candidate variable *relative to the implicit modal context*, and *R* takes privilege as the proportional cause.

In general, two variables are in proper competition with each other over which is proportional to some effect variable only when they are exhaustive relative to the same modal context. *C* and *R* are not competitors for proportionality relative to *Y*, since only one of them can contain an exhaustive set of active possibilities relative to any given modal context.

6. Preserving Causal Intuitions

The strongest objection to proportionality, as raised by Bontly, Shapiro and Sober, McDonnell, and Weslake, is that it seems to render many common sense causal claims false.[12] Call this the *objection from common sense*. It objects to strong proportionality by attempting to demonstrate that if proportionality is required of something to be a cause, then many things that we would naturally call causes don't actually qualify.

Take as an example the situation where Socrates drinks hemlock and then dies, and the corresponding causal claim, 'Socrates's drinking hemlock caused him to die'. The objection goes that drinking hemlock is not actually proportional to Socrates dying. For example, if Socrates had not drank hemlock, but still consumed it – by eating a dozen leaves, for example – then he still would have

[12] See Bontly 2005; Shapiro and Sober 2012; McDonnell 2017; and Weslake 2013, 2017

died. This seems to show that the changes in the variable that represents Socrates drinking hemlock don't line up with the changes in the variable that represents Socrates dying. The first variable could change values from *Socrates-drinks-hemlock* to *Socrates-eats-hemlock* and the second variable would retain the value *Socrates-dies*. This common sense causal claim is therefore not proportional. The proportional cause should be, instead, *consuming hemlock*.

However, this objection is mistaken. It fails to respect the exhaustivity constraint on variable selection, and thereby equivocates between different background modal contexts. It further fails to respect exclusivity, and thereby runs together what should be different variables. Rectifying this illuminates the implicit proportionality of common sense causal claims.

First, the objection ignores the fact that proportionality, in requiring exhaustive and exclusive variables, is relative to modal context. Take the hemlock example just outlined. Importantly, this example and corresponding claim are under-defined.[13] Translated into interventionist terms, all that this description provides is that there is some variable that takes a value that represents Socrates drinking hemlock, and an intervention on this variable changes the value of some other variable to one that represents Socrates dying. But, a number of different variables could represent the purported cause, and a number of different models could represent its relationship to the effect of Socrates' dying. Which of these is accurate depends on what the relevant alternatives to drinking hemlock are. How these details get filled in will determine whether or not the variable that represents Socrates drinking hemlock is proportional.

I hold that the common sense claim that drinking hemlock causes Socrates's death implicitly takes the relevant alternative to be Socrates's *not* drinking hemlock. The default context is taken to be that hemlock was the only possible poison, and drinking it the only possible means of consumption. Given this context, the exhaustive variable would take the values {*drinks-hemlock*, *doesn't-*

[13] I take this to be common knowledge. See Franklin-Hall 2016; McDonnell 2017; and Weslake 2017

drink-hemlock}. But, such a variable is indeed proportional to the effect variable. Thus, the common sense cause is, in fact, proportional.

Such a defense requires that common sense claims be implicitly relative to a modal context. I'm not the first to relativize common sense claims to context. Philosophers such as Mackie and Schaffer make such a move, albeit with different ends in mind.[14] However, both McDonnell and Weslake explicitly deny this kind of relativity.[15] They claim that the very fact that we have strong and convergent intuitions about common sense examples, despite their being under-determined, demonstrates that the intuitions are not sensitive to filling in details.

In response, I argue that we respond to common sense causal examples in the same way that we respond to standard conversations. According to Grice, communication is governed by a set of conversational maxims.[16, 17] The maxims most relevant to how an audience engages with these under-defined causal examples are the maxims of *quantity* and *relation*. Taken together, these maxims enjoin an interlocutor to,

> Make your contribution as informative as is required (for the current purposes of exchange)….[and no] more informative than is required,….[and b]e relevant. (1989, 26 – 27)

Thus, the conversationally natural way to fill in the modal context of these examples is to take each fact as informative and relevant, and to assume that all informative facts have been provided.

The only information provided by the hemlock example is the following: (i) Socrates drinks hemlock. (ii) Socrates dies. The Gricean maxims tell us that this is all the information needed, and that nothing significant has been left out. So, the details are filled in as continuous with everyday life. In possible world speak,

[14] See Mackie 1974, especially chapter 2; and Schaffer 2005
[15] McDonnell 2017; Weslake 2017
[16] See Grice 1989
[17] Bontly makes a similar point (see 2005)

we're looking only at worlds which have a similar environment, a biologically similar Socrates, etc., and in which laws of metaphysical necessity hold.

The causal focus is on Socrates's drinking hemlock. This means that in evaluating the causal relationship, everything else is held fixed and the fact of the drinking hemlock is varied. Due to the absence of any other details, the only real alternative to Socrates's drinking hemlock is his not drinking hemlock. Nothing suggests that there are alternative means of consuming the hemlock. Further, it's not a common occurrence in everyday life to have alternative means of consuming a given poison. Treating *eating hemlock* as a relevant alternative would be to arbitrarily introduce something that wasn't otherwise specified, and whose presence can't be justified by everyday experience.

The objection from common sense assumes different possible alternatives than what I take to be implicit, and then tries to say that relative to these other alternatives, the common sense causal claim is not proportional. I have argued that the common sense cause is simply not relative to these other alternatives.

However, even given other possible alternatives, the common sense cause would still be proportional. The second mistake that the objection makes is that it fails to appreciate the constraint of exclusivity.

The objection holds that there is some relevant alternative to Socrates's drinking hemlock that preserves his consuming it. Take as an arbitrary alternative his eating hemlock. Socrates could both drink and eat the hemlock – he could wash down a hemlock salad with a glass of hemlock milk, for example. Following exclusivity, then, these possibilities should be represented by distinct variables – one that can take the value *drinks-hemlock*, call this *D*, and one that can take *eats-hemlock*, call this *E*.

But, now there is no problem. Following Woodward's response to early pre-emption cases,[18] we can hold *E* fixed at the value that represents Socrates not eating the hemlock, and see if the changes in *D* – which we can ensure meets exhaustivity by giving it the second value *doesn't-drink-hemlock* – line up with the changes in the effect variable. They do. When an intervention sets the value of the cause variable to *drinks-hemlock*, the effect variable takes the value *dies*. When an intervention sets the value of the cause variable instead to *doesn't-drink-hemlock*, the effect variable changes value to *doesn't-die*. Once again, the common sense cause is proportional.

If, on the other hand, the situation is such that Socrates's drinking hemlock is indeed mutually exclusive with his eating hemlock, then *drinks-hemlock* and *eats-hemlock* could be values of the same variable. Imagine that Socrates's jailor only has enough money to purchase either hemlock leaves or hemlock milk, but not both. In this case, neither Socrates's drinking nor his eating will be proportional. The proportional cause is instead his consuming hemlock. The proportional variable will therefore be one that takes as values {*consumes-hemlock*, *doesn't-consume-hemlock*}.

But, this is not in conflict with common sense – so long as we abstract away from normal everyday circumstances, and instead genuinely fix the situation as one in which Socrates is forced to consume hemlock, arbitrarily receiving hemlock leaves or milk. When, given this background, we're asked what causes Socrates's death, it is natural to say that it was his consuming hemlock. After all, it isn't the drinking nor the eating that makes a difference to whether Socrates dies, since had he not done one he would have done the other. It is his consuming hemlock rather than not.

Finally, I'd like to point out that the intuition that Socrates's consuming hemlock is the more proportional cause is actually misguided. The naïve intuition holds that an exhaustive and exclusive variable with the value *consumes-hemlock* – call this H_1 – is more proportional to the exhaustive and exclusive variable with the

[18] See Woodward 2003

value *drinks-hemlock* – call this H_2. But, the modal context to which H_1 will be exhaustive is different than that to which H_2 will be. They're therefore not even in competition for proportionality. Instead, I suggest that this intuition is a response to the fact that H_1's modal context is more inclusive than that of H_2. H_1 can accurately (and proportionally) represent the cause of Socrates's death in a wider range of situations than can H_2. But, this is about stability – as earlier defined – not about proportionality. The model that employs H_1 is simply *more stable* than that which employs H_2. This putative proportionality intuition is actually responding to the property of stability.

7. Conclusion

In this paper, I have defended the interventionist formulation of proportionality by explicating the exhaustivity and exclusivity constraints, and stipulating that proportionality requires variables that meet these constraints.

These constraints have been defined on the assumption that a variable represents a particular object's instantiations of a particular type of property. But, they are easily generalized to cover alternate objects of representation. Take events, for example. If variables represent particular kinds of events occurring or failing to occur, then exhaustivity would require that the values of a variable cover the entire range of possibilities of event occurrence for whatever type of event the variable represents. Exclusivity would require that the values of a variable be event occurrences such that no two could occur simultaneously.

Finally, I have articulated how the interventionist formulation of proportionality responds to the objection from common sense. Such an objection dissolves once the explicated constraints on variable selection are honored.

8. References

Bontly, Thomas. 2005. "Proportionality, Causation, and Exclusion." *Philosophia* 32 (1-4): 331 – 48

Franklin-Hall, Laura. 2016. "High-Level Explanation and the Interventionist's 'Variables Problem.'" *British Journal for the Philosophy of Science* 67 (2):553 – 77

Greenland, Sander, and Babette Brumback. 2002. "An Overview of Relations Among Causal Modelling Methods." *International Journal of Epidemiology* 31:1030 – 37

Grice, H. Paul. 1989. *Studies in the Way of Words.* Cambridge, MA: Harvard University Press

Hitchcock, Christopher. 1996. "The Role of Contrast in Causal and Explanatory Claims." *Synthese* 107 (3): 395 – 419

--- 2009. "Causal Modelling." In *The Oxford Handbook of Causation*, ed. Helen Beebee, Christopher Hitchcock, and Peter Menzies, 299 – 314. Oxford: Oxford University Press

Lewis, David. 1983 "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343 – 77

List, Christian, and Peter Menzies. 2009. "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106 (9): 475 – 502

Mackie, J. L. 1974. *The Cement of the Universe.* Oxford: Oxford University Press

McDonnell, Neil. 2017. "Causal Exclusion and the Limits of Proportionality." *Philosophical Studies* 174 (6): 1459 – 74

Menzies, Peter. 1996. "Probabilistic Causation and the Pre-emption Problem." *Mind* 105 (417): 85 – 117

Menzies, Peter, and Christian List. 2010. "The Causal Autonomy of the Special Sciences." In *Emergence in Mind*, ed. Cynthia Macdonald and Graham Macdonald, 108 – 28. Oxford: Oxford University Press

Papineau, David. 2013. "Causation is Macroscopic but not Irreducible." In *Mental Causation and Ontology*, ed. Sophie C. Gibb and Rögnvaldur Ingthorsson, 126 – 52. Oxford: Oxford University Press

Paul, L.A. 2000. "Aspect Causation." *The Journal of Philosophy* 97 (4): 235 – 56

Schaffer, Jonathan. 2005. "Contrastive Causation." *The Philosophical Review* 114 (3): 297 – 328

Shapiro, Larry, and Elliott Sober. 2012. "Against Proportionality." *Analysis* 72 (1): 89 – 93

Weslake, Bradley. 2013. "Proportionality, Contrast, and Explanation." *Australasian Journal of Philosophy* 91 (4): 785 – 97

--- 2017. "Difference-Making, Closure, and Exclusion." In *Making a Difference,* ed. Helen Beebee, Christopher Hitchcock, and Huw Price, 215 – 32. New York: Oxford University Press

Woodward, James. 2003. *Making Things Happen.* Oxford: Oxford University Press

--- 2008a. "Mental Causation and Neural Mechanisms." In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, ed. Jakob Hohwy & Jesper Kallestrup, 218 – 62. Oxford: Oxford University Press

--- 2008b. "Response to Strevens." *Philosophy and Phenomenological Research* 78 (1): 193 – 212

--- 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biological Philosophy* 25 (3): 287 – 318

--- 2015. "Interventionism and Causal Exclusion." *Philosophy and Phenomenological Research* 91 (2): 303 – 47

--- 2016. "The Problem of Variable Choice" *Synthese* 193 (4): 1047 – 72

Yablo, Stephen. 1992. "Mental Causation" The Philosophical Review 101 (2): 245 – 80

Species as Models

Abstract: This paper argues that biological species should be construed as abstract models, rather than biological or even tangible entities. Various (phenetic, cladistic, biological etc.) species concepts are defined as set-theoretic models of formal theories, and their logical connections are illustrated. In this view organisms relate to a species not as instantiations, members, or mereological parts, but rather as phenomena to be represented by the model/species. This sheds new light on the long-standing problems of species and suggests their connection to broader philosophical topics such as model selection, scientific representation, and scientific realism.

1 Introduction

Biological species has arguably been one of the most controversial topics in the philosophy of biology. Philosophers and biologists alike have long debated over "correct" concepts of species and their ontological status. The traditional account took species as a category, class, or type instantiated by individual organisms. After the advent of evolutionary theory, the typological concept came under fire by those who identify species with a part of biological lineage (Ghiselin 1974; Hull 1976). They forcefully

argued that a species is not an abstract type but a concrete historical entity of which individual organisms are mereological bits. Although this individualist thesis became a de-facto standard in the philosophy of biology in the last century, some have complained its lack of explanatory power and called for a revival of a type or natural-kind based concept of biological species (Boyd 1999).

To this debate between individualists and typologists, this paper introduces yet another thesis according to which species taxa are models of scientific theory. Model is a notoriously equivocal concept, but in this paper it is understood as a set-theoretic entity that makes sentences of a given theory true or false. This implies that biological species are mathematical, rather than biological or even tangible, entities. To work out this claim I begin Section 2 with a reconstruction of various (e.g., phenetic, cladistic, biological etc.) species concepts in terms of formal models that licence characteristic sets of inferences. The model-theoretic rendering illustrates logical connections among different species concepts and provides a platform to evaluate them as a problem of *model selection*. Section 3 then expounds on philosophical implications of the model-theoretic interpretation. Identifying species with models entails that the organism-species relationship is not instantial or mereological, but rather representational; i.e., species as models *represent* individual organisms. This opens the possibility of applying general philosophical discussions on scientific representation and realism to vexed questions concerning the epistemic and ontological status of biological species. Through these arguments this paper puts the species problem under broader contexts of model selection, scientific representation, and scientific realism, depicting it as a special case of the generic question as to how science investigates the world.

2 Species as models

This section fleshes out the main claim of this paper by reconstructing various species concepts as set-theoretic models. The central idea is that species concepts specify theories that underpin biological inferences and descriptions, and species are models that satisfy such theories.

2.1 Typological species concepts

The traditional typological view defines species by its essence, or necessary and sufficient conditions or traits. This finds a straightforward expression as a biconditional form $\forall x(Sx \leftrightarrow T_1x \wedge T_2x \wedge \cdots)$. The extension of species S that satisfies this formula then is the intersection $\bigcap_i \mathbf{T}_i$ (see Figure 1(a)).

Though crude as it is, the biconditional formulation allows certain inferences from traits to species and vice versa. It is this kind of logical reasoning that has enabled, for example, the famous French zoologist George Cuvier to reconstruct the anatomy of a whole organism from just a single piece of bone. As is well known, however, such inferences have very restricted validity, because in most cases it is impossible to find a definite set of phenotypic or genetic characteristics that exclusively defines a given species. Evolution implies species boundaries to be necessarily "fuzzy," which undermines simple biconditional forms. The typological species concept has thus been criticized for its lack of expression ability: a simple algebra of trait-sets cannot capture the nuanced reality of biological species.

(a) Typological Sets (b) Cluster Phenotypic space (c) Cladistic Posets (d) Relational Graphs

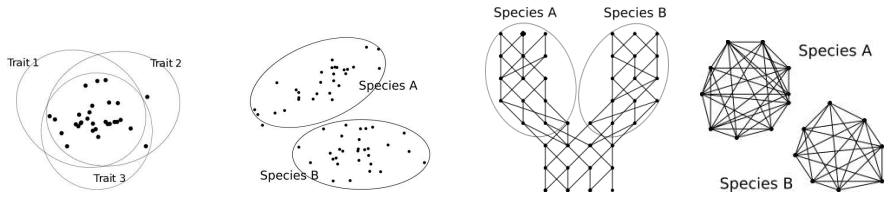


Figure 1: Illustrations of models of various species concepts, with corresponding formal setups. In each model dots/nodes represent individuals. See text for explanation.

2.2 Cluster species concepts

The cluster species concepts avoid this difficulty by defining a species as a group or cluster of similar organisms that do not necessarily share a common set of traits. The question then is how to define similarity. Its earliest variant, the phenetic species concept, represents organisms in a multi-dimensional space each axis of which defines a recorded trait (Sokal and Sneath 1963). Phenotypic similarity is then measured by the euclidean distance between two points/organisms, and a chunk or cluster of organisms in this euclidean space is identified as a species (Figure 1(b)). The choice of euclidean distance is not obligatory. One could, for example, measure similarity by the cosine between two points in the normalized phenotypic space, in which case the similarity amounts to correlation, with a species being identified as a correlated cluster or more generally a *probability distribution* over the phenotypic space (Boyd 1999).

The phenotypic space with a certain metric or probability distribution is certainly a much richer machinery than overlapping sets and allows for more nuanced expressions and inferences. The sophisticated theoretical background (euclidean geometry or

probability theory) enables one to measure the similarity among organisms and to make a trait-species inference in the absence of necessarily or sufficient criteria. To what extent such clustering and inference reflect objective species boundaries, however, was disputed, for the similarity calculation depends much on which phenotypic characters are taken into account. It should also be noted that, like the typological concept, the cluster concepts are purely static and lack a means to express the evolutionary past, the point often criticized by more historical approaches to species.

2.3 Cladistic species concepts

The cladistic species concepts focus on evolutionary history and define species solely in terms of phylogenetic relationships, as a "branch" (monophyletic group) in the evolutionary tree (Hennig 1966). Since ancestral relationship is antisymmetric and transitive, phylogeny forms a (strict) *partially ordered set* or *poset* $(\Omega, \prec)$, with Ω corresponding to a set of organism and $\prec$ meaning "is an ancestor of." A cladistic species is then defined as descendants from some founder organism(s) ω_f:

$$\{\omega \in \Omega : \omega_f \prec \omega\}. \tag{1}$$

An obvious advantage of the cladistic concepts is that it is faithful to the fact of evolution, and for this reason it has been most well received by biologists and philosophers alike. It is not, however, without flaws. For one, although the requirement of monophyly specifies a necessarily condition, it is silent as to how big a branch must be to qualify as a species (for even a small family can satisfy (1)), and so far no satisfactory sufficient condition was given (Velasco 2008). The monophyly requirement has also been

criticized to be too strong, for it would count birds as reptiles because the smallest monophylic group including lizards, snakes, and crocodiles also includes birds. That is, the cladistic species concepts make paraphyletic groups like reptilia *meaningless* (*Sensu* Narens 2007), which strikes some to be too high a price to pay.

2.4 Relational species concepts

Another popular approach is to define a species as a group of individuals in a certain relationship to each other. The biological species concept, for instance, defines species as "groups of interbreeding populations that are reproductively isolated form other such groups (Mayr 1942)" so that the required relationship here is mutual crossability. Other variants focus on reproductive competition (Ghiselin 1974) or organisms' capacity to recognize each other as a possible mate (Paterson 1985). All these proposals try to reduce species into mutual relationships (interbreeding, competition, recognition, etc.) between a pair of organisms. If we represent such relationships by an edge between nodes/organisms, a relational species can be defined as an isolated complete subgraph or *clique* in an undirected graph, that is, a group of nodes in which every two distinct nodes are connected but none is connected to outside (Figure 1(d)). Relational species thus find their model in graph theory, where edges represent the relation in question.

A common criticism of relational species concepts is that the focal relationship such as crossability sometimes fails to induce isolated cliques because some organisms at a species boundary can often mate with organisms that are thought to belong another species (e.g. ring species). Moreover, the biological species concept has been criticized to imply every asexually reproducing organism forms a distinct species (for any singleton

node is complete). These criticisms suggest that the real biological network is so "messy" that just a single relationship cannot divide it into distinct cliques in a non-trivial way.

2.5 "Combo" solutions

The model-theoretic rendering makes explicit what each species concept can and cannot meaningfully say about the biological world. Given that most of the criticisms we have seen concern the "cannot say" part, one way to deal with these difficulties is to combine different theories to obtain more complex definitions of species.

For instance, one may combine the cluster and cladistic species concepts and define a species as a *lineage that shares the same or similar phenotypic distribution*:

$$\{\omega \in \Omega : \omega_f \prec \omega \wedge \theta(\omega_f) = \theta(\omega)\} \tag{2}$$

where $\theta : \Omega \to \mathbb{R}^n$ assigns distribution parameters to each organism $\omega \in \Omega$.[1] On this definition one may meaningfully define paraphyletic species and distinguish birds from other reptiles on the basis of the difference in their phenotypic or genetic profiles. It can also account for anagenesis (speciation without branching) and continuity of species between a cladogenesis (splitting event).

If one replaces θ in (2) with a different function $\nu : \Omega \to N$ that maps organisms $\omega \in \Omega$ to their *niche* $\nu(\omega) \in N$, it becomes the *ecological species concept* which defines a species as "a lineage ... which occupies an adaptive zone minimally different from that of

[1]For non-parametric cases, we can set $\theta : \Omega \to \mathbb{R}^\infty$ and modify the definition as $\{\omega \in \Omega : \omega_f \prec \omega \wedge D(\theta(\omega_f), \theta(\omega)) < k\}$ where $D(\bullet)$ is a divergence measure (such as the Kullback-Leibler divergence) and k is a constant.

any other lineage in its range (Van Valen 1976, 233)."

Yet another combination is that of the cladistic and biological species concepts, which would define a species as a maximum monophylic lineage that can mutually interbreed, so that

$$\{\omega_x, \omega_y \in \Omega : \omega_f \prec \omega_x \wedge \omega_f \prec \omega_y \wedge \omega_x \sim \omega_y\} \tag{3}$$

where $\sim$ stands for crossability.[2] This will make up for the lack of a sufficient condition in the cladistic species concept, and accord well with the so-called *evolutionary species concept* which emphasizes the unique "evolutionary tendencies and historical fate" of each species (Wiley 1978, 17). It should be noted that this could also avoid the problem of ring species because two crossable organisms may not necessary share the same ancestor.

2.6 The scientific species problem as a problem of theory choice

The above discussion shows that (i) major species concepts can be defined as models of formal theories, and that (ii) more complex concepts can be obtained by combining basic ones. The model-theoretic approach characterizes each species concept with the formal apparatus it assumes, which in turn determines its expressive power or what can meaningfully be stated about organisms and/or their history (Narens 2007). In general, a richer theoretical apparatus allows for more nuanced expressions, which makes it less liable to counterexamples. This is illustrated in the progression from the typological to

[2]As in the case of the biological species concept, the crossability here must take into account the existence of two sexes.

cluster and then to cluster-cladistic concepts, where in each step the species concept acquires the ability to deal with fuzzy boundaries and evolutionary history, respectively.

It does not necessarily follow, however, that a richer concept is always desirable, because it tends to have a greater degree of freedom and requires more data in actual application. While only phylogenetic information suffices to demarcate cladistic species, the cluster-cladistic concept also requires phenotypic or ecological information, which in many cases may not be available. A stronger semantic power thus comes with a higher epistemic cost, as is often emphasized by pheneticists or cladists in their respective advocacy of the phenotypic cluster and cladistic species concepts.

This suggests that the competition among various species concepts should be understood as a problem of model selection, where different models are evaluated on the basis of their explanatory or descriptive power versus parsimony or operationality (Sober 2008). Indeed, most disputes among advocates of different species concepts arise from their differential emphasis on what aspects of the biological world a disable species concept needs and needs not take into account (Ereshefsky 2001), but the difficulty is that these emphases are often implicit and incommensurable. Although the model-theoretic approach does not arbitrate these debates, it provides a common formal framework that makes explicit the explanatory power and operationality of species concepts and facilitates evaluation of their respective advantage.

3 Philosophical implications

3.1 Species are models

Upon the model-theoretic reconstruction of various species concepts, we now turn to the philosophical thesis that species taxa should be construed as models proposed above, i.e., as set-theoretic entities. To proceed, let me first begin with an analogy from classical mechanics. Classical mechanics is a theory about Newtonian particles, which are customary defined as volumeless points or vectors in a three-dimensional Euclidean space. Newton's celebrated laws like $\mathbf{F} = m\mathbf{a}$ describe temporal evolution of a system composed of such "particles." This system is to be distinguished from any actual physical systems, say the solar system, for one thing, no concrete bodies are volumeless, nor do they indefinitely continue rectilinear motion as prescribed by Newton's first law. Newton's theory, or any other physical theories for that matter, is a description of idealized and abstracted models and not of actual phenomena (Cartwright 1983). That is, models of classical mechanics — which make its laws and statements true — are not concrete, physical entities, but rather abstract mathematical objects that can be constructed within set theory (McKinsey et al. 1953).

The role of models in science has been emphasized by the so-called semantic or model-based view of scientific theories (e.g. van Fraassen 1980; Suppe 1989).[3] In the traditional, logical-positivist view, a scientific theory was supposed to directly describe

[3]This label ("the semantic view") has been used to describe different, and logically independent, theses. In particular, while some philosophers (e.g. Suppes 2002) take a scientific theory as a *description* of models, others *identify* it with a set of models (van Fraassen 1980). In this paper I adopt the former thesis without committing to the latter.

observed data. This has set for positivists the difficult task of reducing theoretical concepts that seemingly lack direct empirical contents to observation vocabulary by way of *bridge laws* or *partial interpretations*. To avoid this difficulty, proponents of the model-based view take a model, rather than observation, as the primary descriptive target of a scientific theory. In this view, a theory specifies an abstract model that idealizes and extracts just salient factors, and only indirectly relates to actual phenomena via such an model.

I submit that the species problem is a variant of the positivist conundrum. Species is a highly theoretical concept, and various proposal of "species concepts" in the past can be understood as attempts to build bridge laws for reducing it to a set of observational or operational criteria. To date more than a dozen of different concepts have been proposed[4], with no general consensus — each has its own strength, but also weakness and exceptions when applied to the rich and heterogeneous biological world. The assumption has been that a species concept must be a faithful description of *actual* biological features or phenomena. But what if this assumption is untenable, or at least unreasonable? The model-based view has been quite popular among philosophers of biology (e.g. Beatty 1981; Lloyd 1988). If we adopt this view and construe evolutionary theory as describing models, then species too must be defined accordingly, i.e., as (a part of) abstract models that satisfy descriptions and/or inferences of the corresponding theory.

What, then, are theories about species? Without claiming to be exhaustive, this paper adopts Suppes's (2002) thesis that a scientific theory must be defined as a set-theoretical predicate. The foremost advantage of this approach is that it enables one

[4]Mayden (1997), for example, counts at least 22 concepts of species.

to easily harness a theory with mathematical apparatus necessary for sophisticated reasoning. As discussed above, contemporary studies on species rely heavily on quantitative methods to calculate similarity or reconstruct a phylogenetic tree from phenotypic or genetic data. Given that such mathematical reasoning requires matching formal models of calculus or probability theory, the straightforward way to define a species is to build it upon these mathematical backgrounds as an extension of these formal models. Section 2 is a preliminary sketch of applying this Suppesian program to various species concepts. If this attempt turns out to be successful, biological species are to be understood as parts of set-theoretic structures, just like Newtonian particles. That is, they are mathematical and abstract constructs, rather than physical or biological entities.[5]

The purpose of the set-theoretic exposition is not just to accommodate quantitative reasoning. Even with less quantitative cases like the biological species concept, it makes implicit assumptions explicit and suggests a way to deal with counterexamples. The problem of ring species, for example, arises from a conflict between the presumption that each biological species must be isolated and the fact that crossability is not necessarily transitive and thus fails to induce equivalence classes. One possible response to this charge then would be to weaken the former assumption and redefine a species just as a (not necessarily isolated) clique in the reproductive network. Clarification of theoretical assumptions helps us to assess other species concepts as well. For example, the phenetic species concept is often claimed to be "theory-free" in that it does not depend on any evolutionary hypothesis. But as we have seen in Sec. 2.2, the calculation of phenotypic

[5]Hence the present thesis should not be confused with the view that species are sets or collections of *organisms* (Kitcher 1984), which, after all, are concrete biological entities.

similarity presupposes a phenotypic space equipped with a particular (e.g., euclidean) metric, which is a fairly strong theoretical assumption. Also, cladists often stress the simplicity and purity of their monophylic species definition that only considers phylogenetic relationships. But in order to make use of likelihood methods to infer such relationships, as is common in practice, a simple poset is not enough: one also needs to assume some genetic or phenotypic distribution, and then there is no in-principle reason to exclude non-monophylic taxa from the definition of species (as (2) in Sec. 2.5).

The final but not least merit of the set-theoretic approach is its flexibility: it allows for a construction of a new species concept by combining existing ones (Sec. 2.5) or adding new theoretical assumptions. For instance, it is common in experimental biology to characterize a species by shared developmental or causal mechanisms: developmental biologists often talk about "the development of the chicken" and medical doctors rely on causal extrapolation when they prescribe a clinically-tested drug for their patient. Such a "causal species" may be defined by isomorphic *causal models*, which combine a probabilistic distribution and a causal graph over variables. Hence the discussion in Section 2 covers just a few samples that can be constructed within this general framework. This does not of course mean that every possible species concept can and must be formalized, but does suggest the potential of the set-theoretic approach to accommodate the use of existing species concepts and to develop novel ones.

3.2 Philosophical implications

Identifying species with theoretical models sheds new light on some vexed philosophical issues, one amongst which concerns how individual organisms are related to species taxa.

Philosophers have long debated whether the organism-species relationship is instantial (organisms are particular *instances* of a species *qua* class), membership (they are *members* of a species *qua* set; Kitcher 1984), or mereological (they are *parts* of a species *qua* genealogical entity; Ghiselin 1997). The model-theoretic approach suggests an alternative account, according to which a species *represents* (a group of) individual organisms. Just as the Rutherford-Bohr model represents the microscopic structure of atoms, models proposed in Section 2 represent biological populations: for example, nodes and edges consisting of the biological species model in Figure 1(d) respectively represent organisms and crossability. Representation captures our intuitive notion that a model and its target phenomenon share salient static or dynamic features up to a certain precision. Given that said, it must be admitted that the criteria and nature of scientific representation are diversified and still open questions (Frigg and Nguyen 2016). Hence calling the species-organism relationship representational does not necessarily demystify it, but at least implies that the problem is not endemic to evolutionary theory: it is rather a version of a broader philosophical issue as to how the use of scientific models help us understanding the world. This means that the arsenal of this rich philosophical literature can and should be consulted to elucidate the nature of the species-organism relationship. Another, more immediate implication is that the membership and mereological accounts must be both abandoned, for whatever the relationship between a model and phenomena turns out to be, the latter must certainly not be a member or part of the former.

Neither is representation identity or instantiation. Ideal gas is not identical to any actual gas, but only approximates thermodynamic characteristics of some. Hence strictly speaking it has no instantiation, but this does not detract its epistemic validity. Likewise

species concepts, as specifications of ideal models, need not directly apply to actual populations. No wild population big enough to qualify as a species would strictly satisfy the requirement of the biological species concept, because actual mating chance is often hindered by physiological, geographical, and other contingencies. In the same vein, a phenetic or genetic cluster is expected to have outliers when applied to a real population. However, the presence of such exceptions should not immediately invalidate the corresponding species concepts, because the value of a species concept consists less in its universal validity than its epistemic serviceability for inferences and explanations of evolutionary or biological phenomena. These two criteria often conflict: Cartwright (1983) even argues that explanatory theories necessarily distort the reality by idealizing the situation and extracting only relevant features, so that properly speaking they are "lies" by design. Cartwright's examples are physics and economics, but her idea also applies to the present context. The primary function of a species concept is to explain biological phenomena rather than to save them, so that a few discrepancies should not be taken as a falsification.

The conflict between exceptionlessness versus explanatory power also underlies the realism-nominalism debate over species. The proponents of the nominalistic thesis who claim a species to be nothing but a totality of individual organisms have motivated their view by criticizing the realist interpretation of species-as-class for its commitment to the typological thinking and failure to deal with the evident heterogeneity of biological phenomena (e.g. Ghiselin 1997). On the other hand, those who attach weight on the role of species concept in induction and explanation have upheld a realist position and treated species as natural kinds (Boyd 1999). The present thesis offers a third alternative, recognizing the explanatory role of species concept without committing to

the ontologically heavy assumption of natural kinds. As we have seen in Section 2, species as models licence particular sets of inferences. The cluster and typological species/models underpin an expectation that physiological or genetic features found in, say, laboratory animals would also be shared by other individuals of the same species, while the evolutionary species concept explains the reason of such intra-specific similarities. These explanations are effectuated by the same model representing numerically distinct individuals or phenomena to be explained. Note that this procedure no more presupposes the existence of the model as an independent, real entity, than do explanations based on, say, ideal gas. Indeed, explanations may be based on fictional models, as is the case with the Ising model in statistical mechanics.

This does not of course mean that models *must be* fictions, or that species do not exist. Recent advocates of scientific realism argue that successful scientific models capture some, especially structural, aspect of reality (Ladyman 2016). Given its affinity to the model-based view of scientific theories, species realists may well apply this line of reasoning to the present context, taking the set-theoretic structures as discussed in Section 2 as representing the reality or "essential feature" of biological species. Whether and to what extent such an argument carry over, however, remain to be examined by a further study.

4 Conclusion

The past debates over biological species have been based on the assumption that species concepts must describe actual biological phenomena, the strict adherence to which tends to rule out all but cladistic species as typological or inexact. The present paper

challenged this assumption and argued that the primary referent of a species concept is a (set-theoretic) model that licences a certain set of inferences specified by the concept. The model-theoretic rendering articulates explanatory power and theoretical assumptions of each species concept and illuminates logical relationships among them. Once species are specified as models, the long-standing competition among different species concepts reduces to a common problem of model selection. This suggests that evaluation of relative merits and demerits of species concepts must be based more on their explanatory power than on exceptionlessness.

On the philosophical side, the shift in the ontological status of species means that the organism-species relationship is not that of instantiation, membership, or mereology, but rather representation. The vexed issue that has troubled philosophers for decades, therefore, boils down to the broader problem as to how and why scientific models can be used to represent and explain the world. This suggests the possibility to apply the rich literature on scientific representation and realism to elucidate the epistemological and ontological nature of biological species.

In sum, the take home message of the present paper is that the species problem is not endemic to biology or evolutionary theory, but rather is a variant of general scientific and philosophical issues of model selection, scientific representation, and realism. The purpose of this paper was just to establish such a parallelism: determining its philosophical implications on specific debates such as realism or pluralism concerning biological species will be a task for future studies.

References

Beatty, John. 1981. "What's Wrong with the Received View of Evolutionary Theory?." *PSA 1980* 2: 397–426.

Boyd, Richard N. 1999. "Homeostasis, species, and higher taxa." In *Species: New Interdisciplinary Essays.* ed. Robert A Wilson, 141–158, Cambridge, MA: MIT Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Ereshefsky, Marc. 2001. *The Poverty of the Linnaean Hierarchy.* Cambridge: Cambridge University Press.

van Fraassen, Bas C. 1980. *The Scientific Image.* Oxford: Oxford University Press.

Frigg, Roman, and James Nguyen. 2016. "Scientific Representation." In *The Stanford Encyclopedia of Philosophy.* ed. Edward N Zalta, Metaphysics Research Lab, Stanford University.

Ghiselin, Michael T. 1974. "A Radical Solution to the Species Problem." *Society of Systematic Biologists* 23: 536–544.

——— 1997. *Metaphysics and the Origin of Species.* Albany, NY: State University of New York Press.

Hennig, Willi. 1966. *Phylogenetic Systematics.* Urbana, IL: University of Illinois Press.

Hull, David L. 1976. "Are species really individuals?" *Systematic Zoology* 25: 174–191.

Kitcher, Philip. 1984. "Species." *Philosophy of Science* 51: 308–333.

Ladyman, James. 2016. "Structural Realism." In *The Stanford Encyclopedia of Philosophy.* ed. Edward N Zalta, Metaphysics Research Lab, Stanford University.

Lloyd, Elisabeth A. 1988. *The Structure and Confirmation of Evolutionary Theory.* Princeton, NJ: Princeton University Press.

Mayden, R L. 1997. "A hierarchy of species concepts: the denouement in the saga of the species problem." In *Species The Units of Biodiversity.* ed. M F Claridge, H A Dawah, and M R Wilson, 381–424, London: Chapman & Hall.

Mayr, Ernst. 1942. *Systematics and origin of species.* New York, NY: Columbia University Press.

McKinsey, John C C, Patrick Suppes, and A C Sugar. 1953. "Axiomatic Foundations of Classical Particle Mechanics." *Journal of Rational Mechanics and Analysis* 2: 253–272.

Narens, Louis. 2007. *Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science.* Mahwah, NJ: Lawrence Erlbaum Associates.

Paterson, Hugh E H. 1985. "The Recognition Concept of Species." In *Species and Speciation.* ed. E. S. Vrba, 21–29, Pretoria.

Sober, Elliott. 2008. *Evidence and Evolution.* Cambridge: Cambridge University Press.

Sokal, Robert R, and Peter H A Sneath. 1963. *Principles of Numerical Taxonomy.* San Francisco, CA: W. H. Freeman and Co.

Suppe, Frederick. 1989. *The Semantic Conception of Theories and Scientific Realism.*: University of Illinois Press.

Suppes, Patrick. 2002. *Representation and Invariance of Scientific Structures.* Stanford, CA: CSLI Publication.

Van Valen, Leigh. 1976. "Ecological Species, Multispecies, and Oaks." *Taxon* 25: 233–239.

Velasco, Joel D. 2008. "The internodal species concept: a response to 'The tree, the network, and the species'." *Biological Journal of Linnean Society* 93: 865–869.

Wiley, Edward O. 1978. "The Evolutionary Species Concept Reconsidered." *Systematic Biology* 27: 17–26.

Historical Inductions Meet the Material Theory

by Elay Shech

Oct. 2018
(Pre-conference version)

Forthcoming in *Philosophy of Science*

Acknowledgements: I am indebted to John Norton and Moti Mizrahi for extremely valuable discussion and comments on earlier drafts of this paper. Thank you also to helpful conversation with the audience at the Auburn University Philosophical Society in the Spring of 2018 and participants in Gila Sher's *Truth and Scientific Change* reading group in the Fall of 2017 at the Sidney M. Edelstein Center for History and Philosophy of Science, Technology and Medicine at the Hebrew University of Jerusalem.

Abstract: Historical inductions, viz., the pessimistic meta-induction and the problem of unconceived alternatives, are critically analyzed via John D. Norton's material theory of induction and subsequently rejected as non-cogent arguments. It is suggested that the material theory is amenable to a local version of the pessimistic meta-induction, e.g., in the context of some medical studies.

1. Introduction

My goal is to contribute to a growing literature that is critical of historical inductions such as the pessimistic (meta-)induction (PMI) argument (Poincaré 1952, 160; Putnam 1978, 25; Laudan 1981) and the problem of unconceived alternatives (Stanford 2001, 2006) against scientific realism, concentrating mostly on the former. The PMI can be construed in different ways (Mizrahi 2015, Wray 2015), viz., as a deductive *reductio ad absurdum* (e.g., Psillos 1996, 1999), a counterexample to the no miracles argument and inference to best explanation argument for scientific realism (e.g., Saatsi 2005, Laudan 1981), or, usually, as an inductive argument (e.g., Poincaré 1952, Putnam 1978, Laudan 1981, Rescher 1987). In the following I will argue against the inductive version of PMI—or any construal of the PMI that makes use of historical induction—using John D. Norton's material theory of induction (Norton 2003, Manuscript). The upshot is that one ought to be critical of historical inductions that seem to fit the general form or pattern of a good inductive argument, but may in fact lack inductive warrant and force. Various critiques have been put against the PMI (e.g., Lange 2002, Lewis 2001, Mizrahi 2013), along with some defenses (e.g., Saatsi 2005). In Section 2 I will present the PMI and briefly discuss some criticism in order to place my own analysis in broader context. Section 3 presents the material theory of induction and argues that it dissolves the PMI, while Section 4 extends such claims to the more recent problem of unconceived alternatives. In Section 5 I note that the material theory of induction does leave room for a local version of the PMI, which holds in some

limited domain, such as in relation to certain medical studies (Ruhmkorff 2014). I end in Section 6 with a short conclusion.

2. The (Inductive) Pessimistic (Meta-)Induction

The modern formulation of the PMI is usually attributed to Laudan (1981) who argued that having genuinely referential theoretical and observational terms, or being approximately true, is neither necessary nor sufficient for a theory being explanatory and predictively successful. More generally, Anjan Chakravartty characterizes the argument as follows:

> [PMI can] be described as a two-step worry. First, there is an assertion to the effect that the history of science contains an impressive graveyard of theories that were previously believed [to be true], but subsequently judged to be false . . . Second, there is an induction on the basis of this assertion, whose conclusion is that current theories are likely future occupants of the same graveyard. (Chakravartty 2008, 152)[1]

The PMI then may take the following form:

> **[Inductive Generalization PMI]**
>
> P(i) Past theory 1 was successful but not genuinely referential or approximately true.
> P(ii) Past theory 2 was successful but not genuinely referential or approximately true.
> …
> C) Therefore, current (and perhaps future) theories are successful but (by induction) probably not genuinely referential or approximately true.

Laudan (1981) suggests that the history of science contains a graveyard of theories that were previously believed to be approximately true and genuinely referential, but that subsequently were judged to be false and not to refer. Estimations of the number of such superseded theories have been debated (e.g., Lewis 2001, Wray 2013) and recently Mizrahi (2016) presents evidence that challenges the "history of science as a graveyard of theories" claim. Others voice concerns regarding the period of history of science used in order to extract historical evidence (e.g., Lange 2002, Fahrbach 2011) or the proper unit of analysis, i.e., theories vs. theoretical entity (e.g., Lange 2002, Magnus and Callender 2004). Similarly, Park (2011, 83) and Mizrahi (2013, 3220-3222) have argued that the PMI is fallacious due to cherry-picking data, biased statistics, and non-random sampling.

My own criticism of the inductive PMI comes from a different avenue. I will assume that the anti-realist does have randomly sampled historical evidence from the correct period of history and with the proper unit of analysis (whatever those

[1] cf. Wray (2015, 61).

may be) that is not biased or cherry-picked. Still, on the material theory of induction the PMI will not be a cogent argument. In other words, I aim to identify what I take to be a more fundamental (although not categorically different) problem with the PMI.

3. PMI Meets the Material Theory

3.1 The Material Theory of Induction in a Nutshell

Consider the following formally identical inductive inferences (Norton 2003, 649):

> P1) Some samples of the element bismuth melt at 271 degrees C.
> C1) Therefore, all samples of the element bismuth melt at 271 degrees C.
>
> P2) Some samples of wax melt at 91 degrees C.
> C2) Therefore, all samples of wax melt at 91 degrees C.

What makes the first argument an inductively strong and cogent argument while the second a weak and non-cogent inductive argument? Norton (2003, Manuscript) has argued that formal theories of induction, which provide universal schemas that are meant to identify the inductions that are licit and those that are not, stand against an insurmountable difficulty when facing such a question.[2] Instead, he offers a material account of induction:

> In a material theory, the admissibility of an induction is ultimately traced back to a matter of fact, not to a universal schema. We are licensed to infer from the melting point of some samples of an element to the melting point of all samples by a fact about elements: their samples are generally uniform in their physical properties. … *All inductions ultimately derive their licenses from facts pertinent to the matter of the induction*. (Norton 2003, 650; original emphasis)

Norton calls the local facts that power inductive inferences "material postulates." Material postulates themselves are supported by other instances of induction that are licensed by different material postulates.

3.2 Material Analysis of PMI

Many of the criticism of the inductive PMI discussed above amount to the claim that the universal schema used by the likes of Laudan (1981), namely, (P3) Some A's are B's, (C3) Therefore, all A's are B's, does not apply in the case of the PMI because various criteria needed to implement the scheme, e.g., random sampling, correct historical period, proper unit of analysis, have not been met. What I wish to do here

[2] I will not defend Norton's theory or claims here. He dedicates an entire book to the matter in Norton (Manuscript).

is conduct a material analysis of the PMI. Considering the above presentation of the PMI in its [Inductive Generalization PMI] form we may ask, what powers the inductive inference, i.e., what material postulate licenses the pessimistic conclusion?

In context of the two inductive arguments considered in Section 3.1, we note that there is no material postulate that licenses the inductive inference in the case of wax (P2 too C2) but there is one in the case of bismuth (P1 to C1): Generally, chemical elements are uniform in their physical properties. By analogy, the presumption of the meta-induction is that each historical case study looked at is an instance of the same thing, a discovery of induction in science. If we are to perform the meta-induction then there needs to be something in the background facts that unifies all such inductions, just like the fact chemical elements are generally uniform in their physical properties warrants the inductive inference regarding the melting point of bismuth. Let us consider several options.

First, perhaps the material fact is that most scientists use a common rule or method in constructing or discovering successful theories, something along the lines of Mill's methods of experimental inquiry in his *System of Logic* (1872, Book III, Ch. 7). If so, the properties of the rule would be used to authorize the induction. Is there such a rule, or perhaps, some common scientific method? A glance at the history of science suggests that this is unlikely. Newton's deduction from the phenomena, is very different from Darwin's inference to best explanation, which in turn differs radically from Einstein's thought experiments with lights beams, trains, and elevators.[3] More generally, there seems to be a consensus among historians and philosophers of science that something like "the scientific method" is really more of an umbrella term for very different methods used by scientists to construct and discover theories. After all, novel problems necessitate novels solutions, and the commonality that does arise in different cases, say, attempts to minimize error or to be objective, is not the kind of commonality that we seek in powering the PMI and drawing the pessimistic conclusion. For instance, in his book *Styles of Knowing: A New History of Science from Ancient Times to the Present*, Chungling Kwa (2011) argues that there is no single, fundamental method used in science: "there is not just one form of Western scientific rationality; there are at least six." The framework of six "styles of knowing," includes the deductive, the experimental, the hypothetical-analogical, the taxonomic, the statistical, and the evolutionary style, and is based on Alistair Crombie's (1994) three-volume work *Styles of Scientific Thinking*. Similar, Ian Hacking (also taking lead from Crombie's work) has argued that there are distinct "styles of reasoning" used in science, such as the postulational style, the style of experimental exploration, the style of hypothetical construction of models by analogy, the taxonomic style, the statistical style, the historical derivation of genetic development, and the laboratory style (Hacking 1992). This further

[3] In fact, see Norton (Manuscript, Ch. 8-9) who argues that even in historical cases where the *same* principle is applied by scientists, viz., inference to best explanation, "at best we can find loose similarities that the canonical examples of inference to best explanation share," so that no common rule of the kind needed to power the PMI can be found (Ch. 8, p. 1).

corroborates the idea that scientific methods used for theory construction and discovery, as well as for scientific explanation, are very diverse.

More generally, scientific theories are not kind of things that portray the type of uniformity needed to license inductive inferences on Norton's material theory. Albeit in a different context, a similar point is nicely made by Mizrahi (2013, 3218):

> A uniform—as opposed to diverse—sample might be a sample of, say, copper rods. From a sample of just a few copper rods that are tested for electrical conductivity, it is reasonable to conclude that all copper rods conduct electricity because, if you have seen one or two copper rods, you have seen them all (given their uniform atomic structure). Scientific theories, however, are not as uniform as copper rods. The point, then, is that any sample of theories is not going to be uniform in a way that is required for a "seen one, seen them all" inductive generalization.

Similarly, and second, perhaps there are some facts about investigating scientist themselves, how they work, and/or the problems situations that they work in, which can unify the historical evidence in a way that provides us with the inductive warrant we seek. Maybe such facts will include something about the psychology of scientists: their fastidiousness and fear of error, their facility at jumping to conclusions, or perhaps their curiosity, logic, creativity, skepticism, etc. However, in a similar manner to the search for a common rule used in constructing successful theories, the history of science furnishes us with scientists that are heterogeneous enough in their psychological traits, and work in such varied contexts, so as not to provide us with any was to unify the various historical cases in a way pertinent to licensing the pessimistic inference of the PMI.

Third, perhaps we can circumvent looking to a common rule of constructing or discovering theories, or searching for common traits among scientists, by noting that the follow candidate material postulate would power the PMI:

> MP-PMI: Generally, successful theories are not genuinely referential and/or approximately true.

But how would we establish MP-PMI? One option is to appeal to the PMI itself, but this would either be circular or else push us to look for another material postulate. Another option is just to grant the MP-PMI as a reasonable assumption. Perhaps anti-realists or instrumentalists would think that this is a sensible starting point, but their target realist opponent would surely reject such an assumption as question begging. Last, perchance there is some fact about explanatory and/or predictively successful theories that renders them, generally, not genuinely referential and/or approximately true? Possibly part of the essence of successful theories is to misrepresent the world? To me this seems highly unlikely and at odds with any levelheaded intuition but, in any case, if we could argue that successful theories are essentially inaccurate then we would not need the PMI in the first place!

Fourth, we may want to construe the PMI in its inductive generalization form as a kind of abductive argument with the following type of material postulate:[4]

[Inductive Generalization PMI – Abductive version]

P(i): The success of past theory 1 (constructed using method m) is not best explained by its truth.
P(ii): The success of past theory 2 (constructed using method m) is not best explained by its truth.
…
MP: Scientific theories constructed using method m are generally uniform with respect to what best explains their predictive success.
C: The success of our best current (and perhaps futures) theories (constructed using method m) are not best explained by their truth.

Stating the PMI as above has the merit of directly engaging with the "no miracles argument" for scientific realism, namely:

> That terms in mature scientific theories typically refer [to things in the world] …, that theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when it occurs in different theories—these statements are viewed by the scientific realist not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate scientific description of science and its relations to its objects. (Putnam 1975, 73)

But worries abound. First, the realist may very well deny P(i), P(ii), etc., and argue that the success of past theories is best explained by their truth but that, as it turns out, either the best explanation did not hold in this case or else there is some sense in which past theories, insofar as they were successful, were approximately true or on the road to truth. Second, construing the argument as an abduction opens up a Pandora's box of problems associated with the notion of explanation: What is explanation? Are there accounts of explanation where success is best explained by truth and ones in which it isn't and, if so, which account of explanation is relevant in this context? And so on.

Third, the cogency of the argument depends on the idea that all theories appealed to were constructed with some method m, but we already judged that there is no one method that is relevant to constructing scientific theories. Perhaps phenomenological models are good candidates for the type of things that can provide empirical success but are not generally approximately true.[5] Thus, at best, the above argument can power a kind of local PMI: Successful theories constructed

[4] Thanks to Tim Sundell for suggest this line of thought.
[5] Phenomenological models are, generally, not considered explanatory.

by method m are not approximately true. We'll consider one such case in more detail in Section 5.

In short, on the material theory of induction inductive arguments are powered by facts, by material postulates, but in the context of the PMI it seems unlikely that any such non-question begging postulates, which wouldn't render the PMI obsolete, can be found. This is so even if, say, the historical data was not cherry-picked, and the right unit of analysis and correct period of history were used. In other words, I'm equally skeptic of projects that attempt to block the pessimistic conclusion by, for example, taking a random sample of past scientific theories, e.g., Mizrahi (2016). In the following section I'll attempt to extend such claims to the problem of unconceived alternatives.

4. Extension to the Problem of Unconceived Alternatives

Recently, P. Kyle Stanford (2001, 2006) has developed what may be characterized as a new version of the PMI:

> … I propose the following New Induction over the History of Science: that we have, throughout the history of scientific inquiry and in virtually every field, repeatedly occupied an epistemic position in which we could conceive of only one or a few theories that were well-confirmed by the available evidence, while subsequent history of inquiry has routinely (if not invariably) revealed further, radically distinct alternatives as well-confirmed by the previously available evidence as those we were inclined to accept on the strength of that evidence. (Stanford 2001, S8-S9)

The problem of unconceinved alternatives as an argument against scientific realism has been criticized on various grounds (e.g., Chakravartty 2008, Devitt 2011, Mizrahi 2015), but my goal here is just to note that the discussion of Section 3 can be extended to this new version of the PMI, which can be construed as follows:

> P(i) In the past time of theory 1, theory 1 was successful but there were unconceived alternative theories that were as well supported by available evidence but with radically different ontology.
> P(ii) In the past time of theory 2, theory 2 was successful but there were unconceived alternative theories that were as well supported by available evidence but with radically different ontology.
> …
> C) Therefore, in present times, current theories are successful but (by induction) there probably are unconceived alternative theories that are as well supported by available evidence but with radically different ontology.

What we need for the material analysis is something like: Generally, successful theories are underdetermined by data due to possible unconceived alternative theories. In a similar fashion to the MP-PMI, we could look to some common rule used by scientists to conceive theories, or some common psychological traits among

scientist, that may ground the idea that successful theories are such that empirically adequate unconceived alternatives always exists. But for the same reasons discussed above, it seems unlikely that any such common rule or traits will be found. That said, perhaps cognitive facts about human scientists might support the inductive inference to the conclusion that we always miss some alternative theories, which in turn are consistent with the available evidence. What is attractive about this line of thought is that it does seem plausible that due to our cognitive limitations there are always "unconceived alternatives." However, mere cognitive limitations do not support the further conclusion that there are unconcieved alternative theories that are *consistent with available evidence*.

Alternatively, one may think that Stanford's new induction circumvents the material objection: modal reflections alone convince us that there are always unconceived alternative theories that can explain and predict empirical phenomena just as well or better than conceived theories. But how can we come to such a conclusion based on modal reflections alone? Isn't it conceivable if not possible that there would be a point in history with no unconceived alternatives and isn't conceivable if not possible that we are at such point in time in history? Moreover, it is unclear what to make of theory-independent modal claims (unless one has logical modality in mind, which isn't the case here). Certainly, we can talk about different physically possible worlds given a particular physical theory. For instance, various solutions to the Einstein field equations are taken to denote different possible universes according to relativity theory. But it isn't clear what is meant by different possible or alternative conceivable *theories* given no meta-theory as a constraint, so to speak.[6] In any case, if we know that unconceived alternative theories always exist based on modal reflections alone, then the historical induction is doing no work for us at all.

5. Room for a local, material pessimistic induction?

Although the material analysis given here may prompt us to be skeptical of historical inductions (insofar as one is moved by the material theory of induction), it can help us understand why *local* pessimistic inductions may be tenable. Specifically, I want to look at a recent discussion by Rumkorf (2014) who contends that meta-analyses in medicine such as Ioannidis' (2005a, 2005b), which show that a disconcertingly high percentage of prominent medical research findings are refuted by subsequent research, can be developed into a local pessimistic induction. Ioannidis (2005a, 2005b) is concerned with studies, denoted "M-studies," that satisfy the following criteria: "being highly cited, using contemporary research and statistical methods, and being among the first studies to investigate a question at issue" (Rumkorf 2014, 420). Rumkorf's (2014, 421) then uses the various conclusions of Ioannidis (2005a, 2005b) to generate a local PMI in the field of medicine (PMI-M):

[6] What would count as a (logically possible but physically) impossible theory in such a context?

> E1 41% of the associative or causal claims made by M-studies in the sample were inconsistent with the results of subsequent published studies either (1) because the later studies provided evidence against the existence of the association or effect; or (2) because the later studies provided evidence that the magnitude of the association or effect was significantly different.
>
> E2 Therefore, we can expect approximately 41% of the associative and causal claims made by M-studies to be inconsistent with subsequent published studies.

On Norton's theory we need to appeal to a material postulate to license the pessimistic inductive inference in the transitions from E1 to E2, but since we are now working in a limited domain without many heterogeneous examples as in the whole history of science, we may now find some significant commonality between the methods used in different M-studies that can act as licensing facts. What are the background facts that power the PMI-M? Here are some options extracted from Ioannidis's diagnosis of his meta-analysis and quoted in Ruhmkorf (2014, 219):

> Contributing factors include: bias in research (Ioannidis 2005b); non-randomized trials (Ioannidis 2005a); smaller rather than larger sample sizes in refuted studies (Ioannidis 2005a, 224); and publication and time-lag biases (whereby studies with highly significant and potentially aberrational positive results are overrepresented among published articles in major journals and are published more quickly than other articles) (Ioannidis 2005a, 224). Particularly intriguing is the idea that large-scale features of the structure of medical and biological inquiry contribute to the high contradiction rate. Having a number of distinct working groups looking at the same problem increases the chances that at least one of them will find something statistically significant, especially if they are looking at a wide array of possible relationships (Ioannidis 2005b, 697–698). The computational power and richness of data sets available to researchers increases the chance that some of them will be successful in achieving statistical significance, even when no real relationship exists (Ioannidis 2005b, 701).[7]

These various factors, insofar as they are common to most M-studies, are the type of background facts that warrant the pessimistic induction from a material point of view. One may worry of course that the pessimism associated with local PMI generalizes since, presumably, facts about biases and the like are facts about researchers in general, not just researchers in medical science in particular. But, although all scientific studies have to deal with challenges such bias, it may be the case that a particular local subfield, due to its specific nature and whatever social

[7] It should be noted that there are some problems with Ioannidis's (2005a, 2005b) methodology, as identified in Ruhmkorff (2014, 419-421), but they do not seem to be problematic enough to render the PMI-M not cogent.

norms are in place for collecting and disseminative evidence, is especially challenged in a way that can justify the pessimistic induction. The above suggests that this is indeed the case for M-studies.

To end, Ruhmkorff (2014) argues against global PMI on independent grounds (namely, he argues that the PMI commits a statistical error previously unmentioned in the literature and is self-undermining), and but he also argues for the plausibility of a local PMI, viz., M-PMI, and contends that there are clear advantages of PMI-M over PMI. What I wish to note here is that an additional advantage of PMI-M, or local pessimistic induction generally speaking, is that whereas global PMI dissolves upon a material analysis, a material account of PMI-M does seem viable.

6. Conclusion

I have argued that historical inductions such as the (global) PMI and the problem of unconceived alternatives dissolve if we work with the material theory of induction. The reason is that we lack the material postulates needed to license the pessimistic inference: the great heterogeneity of case studies from the history of science of conceiving, constructing, and discovering (explanatory and predictively successful) theories, along with abundant variety of context that scientists find themselves in and traits that they exhibit, make it unlikely that any commonality will be found strong enough to authorize the induction. One may of course object: so much worse for the material theory of induction! This is a fair point, but there is a more general moral to consider. In various situations one may be able to appeal to the notion of "induction" without much being at stake, but in the context of historical inductions like the PMI and problem of unconceived alternatives "induction" is doing a lot of (philosophically) heavy lifting and so the situation rightful calls for scrutiny. Such scrutiny has led to the various discussed criticism that are presented in the context of more traditional, non-material theories of induction. Accordingly, it seems appropriate to show that—even if we assume randomly sampled historical evidence from the correct period of history and with the proper unit of analysis that is not biased or cherry-picked, with no statistical error, etc.—historical inductions do not fare well on the material side of things. I leave objections to the effect that one ought to construe the PMI as a deductive argument, or through a different framework for induction, e.g., via hypothetical or probabilistic induction, for future work.

References

Chakravartty, A. 2008. "What You Don't Know Can't Hurt You: Realism and the Unconceived." *Philosophical Studies* 137: 149–158.

Crombie, A. C., 1995. *Styles of Scientific Thinking in the European Tradition*, 3 vols. London: Duckworth.

Devitt, M. 2011. "Are Unconceived Alternatives a Problem for Scientific Realism?" *Journal for General Philosophy of Science* 42: 285–293.

Fahrbach, L. 2011. "How the Growth of Science Ends Theory Change." *Synthese* 180: 139–155.

Hacking, I. 1992. "'Style' for historians and philosophers." *Studies in History and Philosophy of Science*, 23(1), 1–20.
Ioannidis, J. P. A. 2005a. "Contradicted and Clinically Stronger Effects in Highly Cited Clinical Research." *Journal of the American Medical Association* 294: 218–228.
Ioannidis, J. P. A. 2005b. "Why Most Published Research Findings Are False."*PLoS Medicine 2*: 696–701.
Kwa, C. 2011. *Styles of Knowing: A New History of Science from Ancient Times to the Present*. Pittsburgh: University of Pittsburgh Press.
Lange, M. 2002. "Baseball, Pessimistic Inductions, and the Turnover Fallacy." *Analysis* 62: 281–285.
Laudan, L. 1981. "A Confutation of Convergent Realism." *Philosophy of Science* 48: 19–49.
Lewis, P. J. 2001. "Why the Pessimistic Induction Is a Fallacy." *Synthese* 129: 371–380.
Magnus, P. D., and C. Callender. 2004. "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71: 320–338.
Mill, J. S. [1872] 1916. *A System of Logic: Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. 8th ed. London: Longman, Green, and Co.
Mizrahi, M. 2013. "The Pessimistic Induction: A Bad Argument Gone too Far." *Synthese* 190:3209–3226.
Mizrahi, M. 2015. "Historical Inductions: New Cherries, Same Old Cherry-picking." *International Studies in the Philosophy of Science* 29: 129–148.
Mizrahi, M. 2016. "The history of Science as a Graveyard of Theories: A Philosophers' Myth?" *International Studies in the Philosophy of Science* 30: 263-278.
Norton, J. D. 2003. "A Material Theory of Induction." Philosophy of Science 70: 647–670.
Norton, J. D. Manuscript. *The Material Theory of Induction*. See http://www.pitt.edu/~jdnorton/papers/material_theory/material.html
Park, S. 2011. "A Confutation of the Pessimistic Induction." *Journal for General Philosophy of Science* 42: 75–84.
Poincaré, H. [1902] 1952. *Science and Hypothesis*. New York: Dover. Originally published as *La science et l'hypothèse*. Paris: Flammarion.
Putnam, H. 1978. Meaning and the Moral Sciences. London: Routledge and Kegan Paul.
Psillos, S.: 1996, 'Scientific Realism and the 'Pessimistic Induction' ', *Philosophy of Science* 63 (Proceedings), S306–S314.
Psillos, S. 1999. Scientific Realism: How Science Tracks Truth. London: Routledge.
Rescher, N. 1987. Scientific Realism: A Critical Reappraisal. Dordrecht: D. Reidel.
Ruhmkorff, S. 2013. "Global and Local Pessimistic Meta-inductions." *International Studies in the Philosophy of Science* 27: 409–428.
Saatsi, J. 2005. "On the Pessimistic Induction and Two Fallacies." *Philosophy of Science* 72: 1088–1098.
Sklar, L. M. (2003). "Dappled theories in a uniform world." *Philosophy of Science*, 70, 424–441.

Stanford, P. K. 2001. "Refusing the Devil's Bargain: What Kind of Underdetermination Should We take Seriously?" *Philosophy of Science* 68 (Proceedings): S1–S12.

Stanford, P. K. 2006. Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford: Oxford University Press.

Wray, K. Brad. 2015. "Pessimistic Inductions: Four Varieties." *International Studies in the Philosophy of Science* 29: 61–73.

To be presented at the *2018 PSA Meeting*:

Can Quantum Thermodynamics Save Time?

Noel Swanson*

Abstract

The *thermal time hypothesis (TTH)* is a proposed solution to the problem of time: every statistical state determines a thermal dynamics according to which it is in equilibrium, and this dynamics is identified as the flow of physical time in generally covariant quantum theories. This paper raises a series of objections to the TTH as developed by Connes and Rovelli (1994). Two technical challenges concern the implementation of the TTH in the classical limit and the relationship between thermal time and proper time. Two more conceptual problems focus on interpreting the flow of time in non-equilibrium states and the lack of gauge invariance.

1 Introduction

In both classical and quantum theories defined on fixed background spacetimes, the physical flow of time is represented in much the same way. Time translations correspond to a continuous 1-parameter subgroup of spacetime symmetries, and the dynamics are implemented either as a parametrized flow on statespace (Schödinger picture) or a parametrized group of automorphisms of the algebra of observables (Heisenberg picture). In generally

*Department of Philosophy, University of Delaware, 24 Kent Way, Newark, DE 19716, USA, nswanson@udel.edu

covariant theories, where diffeomorphisms of the underlying spacetime manifold are treated as gauge symmetries, this picture breaks down. There is no longer a canonical time-translation subgroup at the global level, nor is there a gauge-invariant way to represent dynamics locally in terms of the Schrödinger or Heisenberg pictures. Without a preferred flow on the space of states representing time, the standard way to represent physical change via functions on this space taking on different values at different times, also fails. This is the infamous *problem of time*.

Connes and Rovelli (1994) propose a radical solution to the problem: the flow of time (not just its direction) has a thermodynamic origin. Equilibrium states are usually defined with respect to a background time flow (e.g., dynamical stability and passivity constraints reference a group of time translations). Conversely, given an equilibrium state one can derive the dynamics according to which it is in equilibrium. Rovelli (2011) exploits this converse connection, arguing that in a generally covariant theory, *any* statistical state defines a notion of time according to which it is an equilibrium state. The *thermal time hypothesis (TTH)* identifies this state-dependent thermal time with physical time. Drawing upon tools from Tomita-Takesaki modular theory, Connes and Rovelli demonstrate how the TTH can be rigorously implemented in generally covariant quantum theories.

The idea is an intriguing one that, to date, has received little attention from philosophers.[1] This paper represents a modest initial attempt to sally forth into rich philosophical territory. Its goal is to voice a number of technical and conceptual problems faced by the TTH and to highlight some tools that the view has at its disposal to respond.

2 The Thermal Time Hypothesis

We usually think of theories of mechanics as describing the evolution of states and observables through time. Rovelli (2011) advocates replacing this picture with a more general *timeless* one that conceives of mechanics as describing relative correlations between physical quantities divided into two classes, *partial* and *full* observables. Partial observables are quantities that physical measuring devices can be responsive to, but whose value cannot be predicted

[1] Earman (2002), Earman (2011), and Ruetsche (2014) are notable exceptions. Physicists have been more willing to dive in. Paetz (2010) gives an excellent critical discussion of the many technical challenges faced by the TTH.

given the state alone (e.g., proper time along a worldline). A full observable is understood as a coincidence or correlation of partial observables whose value can be predicted given the state (e.g., proper time along a worldline at the point where it intersects another worldline). Only measurements of full observables can be directly compared to the predictions made by the mechanical theory.

A timeless mechanical system is given by a triple $(\mathcal{C}, \Gamma, f)$. $\mathcal{C}$ is the configuration space of partial observables, q^a. A *motion* of the system is given by an unparametrized curve in $\mathcal{C}$, representing a sequence of correlations between partial observables. The space of motions, Γ is the statespace of the system and is typically presymplectic. The evolution equation is given by $f = 0$, where f is a map $f : \Gamma \times \mathcal{C} \to V$, and V is a vector space. For systems that can be modeled using Hamiltonian mechanics, Γ and f are completely determined by a surface Σ in the cotangent bundle $T^*\mathcal{C}$ (the space of partial observables and their conjugate momenta p_a). This surface is defined by the vanishing of some Hamiltonian function $H : T^*\mathcal{C} \to \mathbb{R}$.

If the system has a preferred external time variable, the Hamiltonian can be decomposed as

$$H = p_t + H_0(q^i, p_i, t) \tag{1}$$

where t is the partial observables in $\mathcal{C}$ that corresponds to time. Generally covariant mechanical systems lack such a canonical decomposition. Although these systems are fundamentally timeless, it is possible for a notion of time to emerge thermodynamically. A closed system left to thermalize will eventually settle into a time-independent equilibrium state. Viewed as part of a definition of equilibrium, this thermalization principle requires an antecedent notion of time. The TTH inverts this definition and use the notion of an equilibrium state to select a partial observable in $\mathcal{C}$ as time.

Three hurdles present themselves. The first is providing a coherent mathematical characterization of equilibrium states. The second is finding a method for extracting information about the associated time flow from a specification of the state. Finally, in order to count as an emergent explanation of time, one has to show that the partial observable selected behaves as a traditional time variable in relevant limits.

For generally covariant quantum theories, Connes and Rovelli (1994) propose a concrete strategy to overcome these hurdles. Minimally, such a theory can be thought as a non-commutative C^*-algebra of diffeomorphism-invariant

observables, $\mathfrak{A}$, along with a set of physically possible states, $\{\phi\}$.[2] Via the Gelfand-Nemark-Segal (GNS) construction, each state determines a concrete Hilbert space representation $(\pi_\phi(\mathfrak{A}), \mathcal{H}_\phi)$, and a corresponding von Neumann algebra $\pi_\phi(\mathfrak{A})''$, defined as the double commutant of $\pi_\phi(\mathfrak{A})$.

Connes and Rovelli first appeal to the well-known *Kubo-Martin-Schwinger (KMS) condition* to characterize equilibrium states. A state, ρ, on a von Neumann algebra, $\mathfrak{M}$, satisfies the KMS condition for inverse temperature $0 < \beta < \infty$ with respect to a 1-parameter group of automorphisms, $\{\alpha_t\}$, if for any $A, B \in \mathfrak{M}$ there exists a complex function $F_{A,B}(z)$, analytic on the strip $\{z \in \mathbb{C} | 0 < \mathrm{Im} z < \beta\}$ and continuous on the boundary of the strip, such that

$$\begin{aligned} F_{A,B}(t) &= \rho(\alpha_t(A)B) \\ F_{A,B}(t + i\beta) &= \rho(B\alpha_t(A)) \end{aligned} \tag{2}$$

for all $t \in \mathbb{R}$. The KMS condition generalizes the idea of an equliibirum state to quantum systems with infinitely many degrees of freedom. KMS states are stable, passive, and invariant under the dynamics, $\{\alpha_t\}$. Moreover in the finite limit, the KMS condition reduces to the standard Gibbs postulate.

Although the KMS condition is framed relative to a chosen background dynamics, according to the main theorem of *Tomita-Takesaki modular theory*, every faithful state determines a canonical 1-parameter group of automorphisms according to which it is a KMS state. Connes and Rovelli go on to identify the flow of time with the flow of this state-dependent *modular automorphism group*.

In the GNS representation $(\pi_\phi(\mathfrak{A}), \mathcal{H}_\phi)$, the defining state, ϕ, is represented by a cyclic vector $\Phi \in \mathcal{H}_\phi$. If ϕ is a *faithful* state (i.e., if $\phi(A^*A) = 0$ entails that $A = 0$) then the vector Φ is also separating. In this setting we can apply the tools of Tomita-Takesaki modular theory. The main theorem asserts the existence of two unique modular invariants, an antiunitary operator, J, and a positive operator, Δ. (Here we will only be concerned with the latter.) The 1-parameter family, $\{\Delta^{is} | s \in \mathbb{R}\}$, forms a strongly continuous unitary group,

$$\sigma_s(A) := \Delta^{is} A \Delta^{-is} \tag{3}$$

for all $A \in \pi(\mathfrak{A})''$, $s \in \mathbb{R}$. The defining state is invariant under the flow of the modular automorphism group, $\phi(\sigma_s(A)) = \phi(A)$. Furthermore, $\phi(\sigma_s(A)B) =$

[2]See Brunetti et al. (2003) for a formal development of this basic idea.

$\phi(B\sigma_{s-i}(A))$. Thus ϕ satisfies the KMS condition relative to $\{\sigma_s\}$ for inverse temperature $\beta = 1$.

For any faithful state, this procedure identifies a partial observable, the thermal time, $t_\phi := s$, parametrizing the flow of the (unbounded) thermal hamiltonian $H_\phi := -\ln \Delta$, which has Φ as an eigenvector with eigenvalue zero. We can then go on to decompose the timeless Hamiltonian $H = p_{t_\phi} + H_\phi$. Associated with any such state, there is a natural "flow of time" according to which the system is in equilibrium. But in what sense does this thermal time flow correspond to various notions of physical time? In particular, how is thermal time related to the proper time measured by a localized observer?

Although they do not establish a general theorem linking thermal time to proper time, Connes and Rovelli do make substantial progress on the third hurdle in one intriguing special case. For a uniformly accelerating, immortal observer in Minkowski spacetime, the region causally connected to her worldline is the *Rindler wedge*. In standard coordinates we can explicitly write the observer's trajectory as

$$\begin{aligned} x^0(\tau) &= a^{-1}\sinh(\tau) \\ x^1(\tau) &= a^{-1}\cosh(\tau) \\ x^2(\tau) &= x^3(\tau) = 0 \end{aligned} \tag{4}$$

where τ is the observer's proper time. The wedge region is defined by the condition $x^1 > |x^0|$. The *Bisognano-Wichmann theorem* then tells us that in the vacuum state, the modular automorphism group for the wedge implements wedge-preserving Lorentz boosts — Δ^{is} is given by the boost $U(s) = e^{2\pi i s K_1}$ (where K_1 is the representation of the generator of an x^1-boost). Since the Lorentz boost $\lambda(a\tau)$ implements a proper time translation along the orbit of an observer with acceleration a, $U(\tau) = e^{ai\tau K_1}$ can be viewed as generating evolution in proper time. Comparing these two operators, we find that proper time is directly proportional to thermal time,

$$s = \frac{2\pi}{a}\tau \tag{5}$$

The Unruh temperature measured by the observer is $T = a/2\pi k_b$ (where k_b is Boltzmann's constant), this leads Connes and Rovelli to propose that the Unruh temperature can be interpreted as the ratio between thermal and proper time. Not only does this relationship hold along the orbits of constant

acceleration, but if an observer constructs global time coordinates for the wedge via the process of Einstein synchronization, this global time continues to coincide with the rescaled thermal time flow.

We can now summarize the main content of the TTH:

Thermal Time Hypothesis (Rovelli-Connes)**.** *In a generally covariant quantum theory, the flow of time is defined by the state-dependent modular automorphism group. The Unruh temperature measured by an accelerating observer represents the ratio between this time and her proper time.*

This is a bold idea with a numerous potential implications for quantum physics and cosmology. Over the next three sections, we will consider a series of technical and conceptual objections to the TTH.

3 Thermal Time and Proper Time

The Bisognano Wichmann theorem only applies to immortal, uniformly accelerating observers in the vacuum state of a quantum field theory in flat spacetime. How can we characterize the relationship between thermal and proper time for a broader, more physically realistic class of observers and theories?

A uniformly accelerating mortal observer has causal access to a different region of Minkowski spacetime, the *doublecone* formed by the intersection of her future lightcone at birth and her past lightcone at death. Because wedges and doublecones can be related by a conformal transformation, in conformally invariant theories, geometric results from wedge algebras can be transferred onto the doubelcone algebras. In the vacuum state of a conformal theory, the doublecone modular automorphism group acts as Hislop-Longo transformations (Hislop and Longo, 1982). Martinetti and Rovelli (2003) use this result to calculate the corresponding relationship between thermal time and proper time for a uniformly accelerating mortal observer:

$$s = \frac{2\pi}{La^2}(\sqrt{1 + a^2L^2} - \cosh a\tau) \tag{6}$$

where L is the observer's lifetime. (The relationship is more complicated in this case due to the fact that proper time is bounded while modular time is unbounded.) For most of the observer's lifespan, s is an approximately constant function of τ, allowing the Unruh temperature to again be interpreted as the local ratio between thermal and proper time.

This is the best we can hope for. Trebels (1997) proves that arbitrary doublecone automorphisms act as local dynamics, only if they act as scaled Hislop-Longo transformations.[3] Of course, if nature is described by a non-conformal theory, then there is no guarantee that the doublecone modular automorphisms will have a suitable geometric interpretation. Saffary (2005) goes further, arguing that they will not have geometric significance in any theory with massive particles. The mathematical results backing this conjecture, however, are only partial.[4]

Attempting to generalize the TTH to cover non-uniform acceleration and non-vacuum states generates further difficulties. Work on the Unruh effect for non-uniformly accelerating observers (e.g., Jian-yang et al. 1995), indicates that such observers feel an acceleration-dependent thermal bath, reflecting the shifting ratio between constant thermal time and acceleration-dependent proper time. The TTH must explain the phenomenological experience of the observer who will presumably age according to her proper time, not the background thermal time flow. On top of this, if the global state is not a vacuum state, then it is not clear that the wedge modular automorphisms will carry a dynamical interpretation at all. The Radon-Nikodym theorem ensures that the action of the modular automorphism group uniquely determines the generating state. If ϕ, ψ are two (faithful, normal) states on a von Neumann algebra $\mathfrak{M}$, then the associated modular automorphism groups σ^t_ϕ, σ^t_ψ differ by a non-trivial inner automorphism, $\sigma^t_\phi(A) = U\sigma^t_\psi(A)U^*$, for all $A \in \mathfrak{M}$, $t \in \mathbb{R}$, so the general wedge dynamics will not be simple rescalings of the vacuum case.

None of these are knockdown objections since so little is known about the geometric action of modular operators apart from the Bisognano-Wichmann theorem and its conformal generalization. But our current ignorance also presents a major challenge. (The situation is even less clear in general curved spacetime settings.) The defender of the TTH has at least four options on

[3]Formally, Trebels requires that local dynamics be continuous 1-parameter groups of automorphisms of the doublecone algebra that preserve subalgebra localization as well as spacelike and timelike relations between interior points. For a detailed discussion of Trebels's results, see Borchers (2000), §3.4.

[4]In the massless case, the modular generators are ordinary differential operators, δ_0, of order 1. In the massive case, it has been conjectured that the modular generators are pseudo-differential operators $\delta_m = \delta_0 + \delta_r$, where the leading term is given by the massless generator δ_0 and δ_r is a pseudo-differential operators of order < 1. This second term is thought to give rise to non-local action without geometric interpretation.

the table.

She can hold out hope for a suitably general dynamical interpretation of modular automorphisms in a wide class of physically significant states. There is some indication that states of compact energy (e.g., states satisfying the Döplicher-Haag-Roberts and Buchholz-Fredenhagen selection criteria) give rise to well-behaved modular structure on wedges. In this case the wedge modular automorphisms can be related to those in the vacuum state by the Radon-Nikodym derivative (Borchers, 2000). The analogous problem for doublecones is still open.

Alternatively, she could reject the idea that the thermal time flow determines the temporal metric directly. Thermal time would only give rise to the order, topological, and group theoretic properties of physical time. Metrical properties would be determined by a completely different set of physical relations. Some support for this idea comes from the justification of the clock hypothesis in general relativity. Rather than stipulating the relationship between proper time, τ, and the length of a timelike curve $||\gamma||$, Fletcher (2013) shows that for any $\epsilon > 0$, there is an idealized lightclock moving along the curve which will measure $||\gamma||$ within ϵ. This justifies the clock hypothesis by linking the metrical properties of spacetime to the readings of tiny lightclocks. If the metrical properties of time experienced by localized observers arises via some physical mechanism akin to light clock synchronization. This would explain why the duration of time felt by the observer matches her proper time and not the geometrical flow of thermal time.

Perhaps motivated by the justification of the clock hypothesis, the defender of the TTH could attempt to argue that the metrical properties of time emerge from modular dynamics in the short distance limit of the theory. If the theory has a well-defined ultraviolet limit, the renormalization group flow should approach a conformal fixed point. Buchholz and Verch (1995) prove that in this limit, the double-cone modular operators act geometrically like wedge operators implementing proper time translations along the observer's worldline. It is unlikely that the physics at this scale would directly impact phenomenology, but the asymptotic connection might turn out to be important for explaining the metrical properties of spacetime (which bigger, more realistic lightclocks measure) as emergent features of some underlying theory of quantum gravity.

A final option would be to go back to the drawing board. Rovelli and Connes briefly note that since the modular automorphisms associated with each (faithful, normal) state of a von Neumann algebra are connected by

inner automorphisms, they all project down onto the same 1-parameter group of outer automorphisms the algebra. The TTH could be revised to claim that this canonical state-independent flow represents the non-metrical flow of physical time. It is not known, however, under what circumstances the outer flow acts in suitably geometric fashion to be interpretable as local dynamics, so it remains to be seen whether or not this is a viable option. The move does have immediate consequences for the global dynamics, however. Since the global algebra is expected to be type I, all modular automorphisms will be inner. As a result the canonical group of outer automorphisms is trivial. At a global level, there is no passage of time. At the local level, time emerges as a consequence of our ignorance of the global state.

4 The Classical Limit

The classical limit presents a different kind of challenge. Conceptually, nothing about the idea that a statistical state selects a preferred thermal time requires that the theory be quantum mechanical. The proposed mechanism for selecting a partial observable using modular theory, however, does appear to rely on the noncommutativity of quantum observables. If we model classical systems using abelian von Neumann algebras, then every state is tracial (i.e., $\phi(AB) = \phi(BA)$), and consequently every associated modular automorphism group acts as the identity, trivializing the thermal time flow. Does the TTH have a classical counterpart, or is quantum mechanics required to save time in a generally covariant setting?

Arguing by analogy with standard quantization procedures, Connes and Rovelli suggest that in the classical limit commutators need to be replaced by Poisson brackets. We begin with an arbitrary statistical state, ρ, represented by a probability distribution over a classical statespace Γ:

$$\int_{\Gamma} dx \ \rho(x) = 1 \tag{7}$$

where $x \in \Gamma$ is a timeless microstate. By analogy with the Gibbs postulate, we can introduce the "thermal Hamiltonian,"

$$H_{\rho} = -\ln \rho \tag{8}$$

With respect to the corresponding Hamiltonian vector field, the evolution of

an arbitrary classical observable, $f \in C^\infty(\Gamma)$, is given by

$$\frac{d}{ds} f = \{-\ln \rho, f\} \tag{9}$$

and $\rho = \exp(-H_\rho)$. With respect to the Poisson bracket structure, the classical algebra of observables is non-abelian. Gallavotti and Pulvirenti (1976) use this non-abelian structure to define an analogue of the KMS condition. Is this connection strong enough to support a version of the TTH in ordinary general relativity? Or does it only serve to aid us in understanding how the thermal time variable behaves in the transition from quantum theory to classical physics?

The difficulty lies in connecting the thermal time flow for an arbitrary statistical state to our ordinary conception of time. In the quantum case this link was provided by the Bisognano-Wichmann theorem, which does not have a classical analogue. The problem is magnified by the lack of a full understanding of statistical mechanics and thermodynamics in curved spacetime. Rovelli has done some preliminary work on developing a full theory of generally covariant thermodynamics based on the foundation supplied by the TTH, including an elegant derivation of the Tolman-Ehrenfest effect, but the field is still young.[5]

Setting aside these broader interpretive challenges for now, an important first step lies in obtaining a better understanding the classical selection procedure outlined above. As it turns out, the commutator-to-Poisson-bracket ansatz is on firmer foundational footing than one might initially suspect. As emphasized by Alfsen and Shultz (1998), non-abelian C^*-algebras have a natural *Lie-Jordan structure*:

$$AB = A \bullet B - i(A \star B) \ , \tag{10}$$

The non-associative Jordan product, $\bullet$, encodes information about the spectra of observables, while the associative Lie product, $\star$, encodes the generating relation between observables and symmetries. The significance of the commutator, is that it defines the canonical Lie product, $A \star B := i/2[A, B]$. Classical mechanical theories formulated on either a symplectic or Poisson manifold have a natural Lie-Jordan structure as well. The standard product of functions defines an associative Jordan product, encoding spectral information, while the Poisson bracket determines the associative Lie product,

[5] See Rovelli and Smerlak (2011).

describing how classical observables generate Hamiltonian vector fields on statespace. Together, this structure is called a *Poisson algebra*. The primary difference between the classical and quantum cases is the associativity/non-associativity of the Jordan product.

These considerations point towards the idea that the appropriate classical analogue of a noncommutative von Neumann algebra, is not a commutative von Neumann algebra, but a Poisson algebra. In this setting, initial strides towards a classical analogue of modular theory have been made by Weinstein (1997). Given any smooth density, μ, on a Possion manifold, Γ, Weinstein defines a corresponding *modular vector field* ϕ_μ given by the operator $\phi_\mu : f \to \mathrm{div}_\mu H_f$ where H_f is the Hamiltonian vector field associated with a classical observable, $f \in C^\infty(\Gamma)$. The antisymmetry of the Poisson bracket entails that the operator ϕ_μ is a vector field on Γ. Weinstein proposes ϕ_μ as the classical analogue of the modular automorphism group. It characterizes the extent to which the Hamiltonian vector fields are divergence free (with respect to the density μ), vanishing iff all Hamiltonian vector fields are divergence free.

We can connect Weinstein's classical modular theory to the TTH. If Γ is a symplectic manifold and we let μ be the density associated with the canonical Liouville volume form, then $\phi_\mu(f) = 0$ for all observables. This reflects the conservation of energy by Hamiltonian flows in symplectic dynamical systems. Given any statistical state, however, we can define an associated density which leads to a nontrivial modular vector field. For any positive function, h, we have

$$\phi_{h\mu} = \phi_\mu + H_{-\ln h} = H_{-\ln h}. \tag{11}$$

Therefore any statistical state, ρ, defines a modular vector field equivalent to the Hamiltonian vector field $H_{-\ln \rho}$ associated with the density $e^{-\ln \rho}\mu$. We immediately recognize $-\ln \rho$ as the thermal Hamiltonian postulated by Connes and Rovelli. Clearly, $e^{is \ln \rho} \rho e^{-is \ln \rho} = \rho$, thus the state is invariant with respect to the flow of $H_{-\ln \rho}$. Additionally, it can be shown that ρ satisfies the KMS condition with respect to these dynamics, hence, from the perspective of the associated time flow ρ resembles an invariant equilibrium state just as in the quantum case.

5 Conceptual Challenges

As we have seen in the previous two sections, the TTH faces a number of technical challenges (some of which look easier to overcome than others). There are, however, several deeper conceptual problems looming in the background which pose a more serious challenge to the viability of the hypothesis. Here, we will discuss two of the most pressing.

The first, which we will call the *generality problem*, draws upon the preceding discussion of the classical limit. While mathematically speaking, Weinstein's modular vector field gives us a method for selecting a canonical thermal time flow in a classical theory, physical speaking, there is no reason why we should view the corresponding thermal time as physical time. As we have seen, any statistical state determines thermal dynamics according to which it is a KMS state, however, if ρ is a non-equilibrium state, the resultant thermal time flow does not align with our ordinary conception of time. By the lights of thermal time, a cube of ice in a cup of hot coffee is an invariant equilibrium state! The same problem arises in the quantum domain — only for states which are true equilibrium states will the thermal time correspond to physical time.

It appears inevitable that the TTH will have to be tempered. Rather than letting any state determine a corresponding flow of thermal time, only certain reference states should be permitted. Apart from the problem of providing an intrinsic, non-dynamical characterization of such states, if a system is not in one of these, it is hard to envision how a counterfactual state of affairs can determine the actual flow of time.[6] This might provide more reasons for the defender of the TTH to explore the state independent, outer modular flow. Alternatively, she could try to argue that local non-equilibrium behavior can be viewed as small fluctuations from some background state. On this approach, the local flow of time in my office according to which the ice

[6] A closely related worry, what we might call the *background-dependence problem*, has been voiced by Earman (2011) and Ruetsche (2014). Their concern is that we can only identify modular automorphisms as dynamics because we already have a rich spatiotemporal geometry in the background. This casts doubt on whether the TTH can provide a coherent definition of time in situations where such structure is absent (as required to solve the full problem of time). This is exacerbated if the TTH is modified in response to the generality problem. Unless the modular automorphism group can always be viewed dynamically, the defender of the TTH will be hard pressed to find constraints capable of separating the dynamical cases from the non-dynamical cases which are independent of all background temporal structure.

melts and the coffee cools is not defined by the thermal state of the ice/coffee system, but the thermal state of some larger enveloping system (the entire universe perhaps). Rovelli (1993) hints in this direction, calculating that in a Friedman-Robertson-Walker universe, the thermal time induced by the equilibrium state of the cosmic microwave background will be proportional to the FRW time. While the connection is intriguing, it seems unlikely that an explanation of this sort will be able to account for the flow of time experienced by localized, mortal observers like us. It would be truly remarkable to discover that our faculties of perception are sensitive to the thermal features of the CMB.

The second problem is the *gauge problem*. The TTH does succeed in providing a means to select a privileged 1-parameter flow on the space of full, gauge invariant observables of a generally covariant theory. What makes this flow interpretable as a *dynamical* flow, however, is its description as a sequence of correlations between partial observables. The difficulty is that these partial observables are not diffeomorphism invariant. Assuming that we treat diffeomorphisms in generally covariant theories as standard gauge symmetries (which is how we got into the problem of time in the first place), then the partial observables are just descriptive fluff. They do not directly represent physical features of our world.

The problem is *not* the resultant timelessness of fundamental physics. The TTH adopts this dramatic conclusion willingly. The problem is that the TTH is supposed to explain how the appearance of time and change emerge from timeless foundations. But the explanation given is couched in gauge-dependent language, and it is not apparent how we can extract a gauge invariant story from it. We can introduce partial observables and use correlations between them to calculate and predict emergent dynamical behavior, but we cannot use these correlations to *explain* that behavior. We lack a gauge invariant picture of generally covariant theories, and the TTH, at least in its present form, does not provide one.

Can a revised TTH give us the explanatory tools needed to understand the flow of time without reference to partial observables, or, does the entire framework of timeless mechanics require us to revise our conception of how ontology, explanation, and gauge symmetries are related?[7] Whether or not

[7]Drifting in the latter direction, Rovelli (2014) suggests that gauge-dependent quantities are more than just mathematical redundancies, "they describe handles through which systems couple: they represent real relational structures to which the experimentalist has access in measurement by supplying one of the relata in the measurement procedure itself."

quantum thermodynamics can save time may rest on the solutions to these new incarnations of vexingly familiar philosophical problems.

References

Alfsen, E. and F. Shultz (1998). Orientation in operator algebras. *Proceedings of the National Academy of Sciences, USA 95*, 6596–6601.

Borchers, H. J. (2000). On revolutionizing quantum field theory with Tomita's modular theory. *Journal of Mathematical Physics 41*(6), 3604–3673.

Brunetti, R., K. Fredenhagen, and R. Verch (2003). The generally covariant locality principle – a new paradigm for local quantum field theory. *Communications in Mathematical Physics 237*, 31–68.

Buchholz, D. and R. Verch (1995). Scaling algebras and renormalization group in algebraic quantum field theory. *Reviews in Mathematical Physics 7*, 1195.

Connes, A. and C. Rovelli (1994). Von Neumann algebra automorphisms and time-thermodynamics relation in generally covariant quantum theories. *Classical and Quantum Gravity 11*(12), 2899.

Earman, J. (2002). Thoroughly modern McTaggart. *Philosopher's Imprint*, 2. http://www.philosophersimprint.org/002003/.

Earman, J. (2011). The Unruh effect for philosophers. *Studies in History and Philosophy of Modern Physics 42*, 81–97.

Fletcher, S. (2013). Light clocks and the clock hypothesis. *Foundations of Physics 43*, 1369–1383.

Gallavotti, G. and M. Pulvirenti (1976). Classical KMS condition and Tomita-Takesaki theory. *Communications in Mathematical Physics 46*, 1–9.

Hislop, P. D. and R. Longo (1982). Modular structure of the local algebras associated with a free massless scalar field theory. *Communications in Mathematical Physics 84*, 71.

Jian-yang, Z., B. Aidong, and Z. Zheng (1995). Rindler effect for a nonuniformly accelerating observer. *International Journal of Theoretical Physics 34*, 2049–2059.

Martinetti, P. and C. Rovelli (2003). Diamond's temperature: Unruh effect for bounded trajectories and thermal time hypothesis. *Classical and Quantum Gravity 20*(22), 4919.

Paetz, T.-T. (2010). An analysis of the 'thermal-time concept' of Connes and Rovelli. Master's thesis, Georg-August-Universität Göttingen.

Rovelli, C. (1993). The statistical state of the universe. *Class.Quant.Grav. 10*, 1567.

Rovelli, C. (2011). Forget time: Essay written for the FQXi contest on the nature of time. *Foundations of Physics*.

Rovelli, C. (2014). Why gauge? *Foundations of Physics 44*(1), 91–104.

Rovelli, C. and M. Smerlak (2011). Thermal time and Tolman–Ehrenfest effect: 'temperature as the speed of time'. *Classical and Quantum Gravity 28*(7), 075007.

Ruetsche, L. (2014). Warming up to thermal the thermal time hypothesis. Quantum Time Conference, University of Pittsburgh, March 28-29.

Saffary, T. (2005). *Modular Action on the Massive Algebra*. Ph. D. thesis, Hamburg.

Trebels, S. (1997). *Über die Geometrische Wirkung Modularer Automoprhismen*. Ph. D. thesis, Göttingen.

Weinstein, A. (1997). The modular automorphism group of a Poisson manifold. *Journal of Geometry and Physics 23*, 379–394.

Neural redundancy and its relation to neural reuse

Abstract

Evidence of the pervasiveness of neural reuse in the human brain has forced a revision of the standard conception of modularity in the cognitive sciences. One persistent line of argument against such revision, however, draws from a large body of experimental literature attesting to the existence of cognitive dissociations. While numerous rejoinders to this argument have been offered over the years, few have grappled seriously with the phenomenon. This paper offers a fresh perspective. It takes the dissociations seriously, on the one hand, while affirming that traditional modularities of mind do not do justice to the evidence of neural reuse, on the other. The key to the puzzle is neural redundancy. The paper offers both a philosophical analysis of the relation between reuse and redundancy, as well as a plausible solution to the problem of dissociations.

1. Introduction

Cognitive science, linguistics and the philosophy of psychology have long been under the spell of "the modularity of mind" (Fodor 1983), or the idea of the mind as a modular system (see e.g. de Almeida and Gleitman 2018). In contemporary psychology, a modular system is generally understood to be "one consisting of functionally specialized subsystems responsible for processing different classes of input (e.g. for vision, hearing, human faces, etc.), or at any rate for handling specific cognitive tasks" (Zerilli 2017a, 231). According to this theory, "human cognition can be decomposed into a number of functionally independent processes, [where] each of these processes operates over a distinct domain of cognitive information" (Bergeron 2007, 176). What makes one process distinguishable from another is its "functional independence, the fact that one can be affected, in part or in totality, without the other being affected, and vice versa" (Bergeron 2007, 176). Furthermore, given that functional processes are realized in the brain, a functionally specialized process is one which presumably occupies a distinctive portion of neural tissue, though not necessarily a small, closely circumscribed and contiguous region. So fruitful and influential has this model been that it is safe to say that in many quarters of the cognitive sciences—and most especially in cognitive psychology, cognitive neuropsychology and evolutionary psychology—modularity is essentially the received view (McGeer 2007; Carruthers 2006; de Almeida and Gleitman 2018).

Developments in cognitive neuroscience over the past thirty years, however, have discomfited the modular account. More evidence than ever before points to the pervasiveness of neural reuse in the human brain—the "redeployment" or "recycling" of neural circuits over widely disparate cognitive domains (Anderson, 2010, 2014; Dehaene, 2005). As the terminology suggests, theories of "re-use" posit the "exaptation" of established and diachronically stable neural circuits over the course of evolution or normal development *without* loss of original function, so that the functional contribution of a circuit is preserved across multiple task domains.[1] As Anderson (2010, 246) explains, "rather than posit a functional architecture for the brain whereby individual regions are dedicated to large-scale cognitive domains like vision, audition, language and the like, neural reuse theories suggest that low-level neural circuits are used and reused for various purposes in different cognitive and task domains." According to the theory, just the same circuits exapted for one purpose can be exapted for another provided sufficient intercircuit pathways exist to allow alternative arrangements of them. Indeed, the same parts put together in *different* ways will yield different functional outcomes, just as "if one puts together the same parts *in the same way* one will get the same functional outcomes" (Anderson 2010, 247, my emphasis). The evidence here converges from heterogeneous sources and research paradigms, including neuroimaging (Anderson 2007a; 2007b; 2007c; 2008), computational (Eliasmith 2015), biobehavioral (Casasanto and Dijkstra 2010) and interference paradigms (Gauthier et al.

[1] This usage of "exaptation" is somewhat misleading, since exaptation usually implies loss of original function (see Godfrey-Smith 2001).

2003), and exempts practically no area of the brain (Leo et al. 2012, 2), including areas long regarded as specialized hubs for certain types of sensory processing, e.g. visual and auditory pathways (Striem-Amit and Amedi 2014). Among other things, this means that one of the hallmark features of a module—its domain specificity (Coltheart 1999)—looks too stringent a requirement to prove useful.[2] For neural reuse demonstrates that any one module will typically be sensitive to *more* than one stimulus, including—most importantly—those channeled along intermodal pathways. Meanwhile efforts to salvage a computational or "software" theory of modularity, which carries no commitments regarding implementation, have met with scepticism (Anderson 2007c; 2010; Anderson & Finlay 2014) if not outright opposition (Zerilli 2017a).[3] And while the brain could still be modular in some other sense, what is clear is that the strict domain-specific variety of modularity can no longer serve as an appropriate benchmark.[4]

And yet there is a persistent line of argument *against* this conclusion which draws from a large body of experimental literature attesting to the existence of cognitive

[2] The sense of domain specificity that is relevant here refers to a module's sensitivity to a restricted class of inputs as defined by a domain of psychology—such as visual, auditory or linguistic information. For discussion of alternative senses, see Barrett and Kurzban (2006) and Prinz (2006).

[3] Though by no means universally (see e.g. Carruthers 2010; Jungé and Dennett 2010).

[4] Nor, for that matter, can its cognate property, informational encapsulation (see below).

dissociations, in which a cognitive ability (say language) is either selectively impaired (linguistic ability is compromised, but no other cognitive ability seems to be materially affected) or selectively spared (general intelligence is compromised, while linguistic abilities function more or less as they should). This literature, most vividly exemplified in lesion studies, is frequently cited in support of classical modularities of mind—be they inspired by the likes of Jerry Fodor (1983), evolutionary psychology (e.g. Cosmides and Tooby 1994; Barrett and Kurzban 2006; Carruthers 2006) or some variation thereof (e.g. ACT-R). While numerous rejoinders to this line of thinking have been offered over the years, few have grappled seriously with the phenomenon, either dismissing the dissociations as noisy, or reasoning from architectural considerations that even nonmodular systems can generate dissociations (Plaut 1995). The aim of this paper is to offer a fresh perspective on this vexed topic. I take the dissociation evidence seriously, on the one hand, while affirming that traditional modularities of mind do not do justice to the evidence of neural reuse, on the other. I do this by invoking neural redundancy, an important feature of cortical design that ensures we have various copies of the same elementary processing units that can be put to alternative (if computationally related) uses in enabling diverse cognitive functions. In the course of the discussion I offer a philosophical explication of the relationship between neural reuse and neural redundancy.

2. What is the Problem? Cognitive Dissociations and Neural Reuse

Let us take an especially contentious question to underscore the nature of the problem we are dealing with and how redundancy might assist in its illumination. The question is this: Does language rely on specialized cognitive and neural machinery, or does it rely on the same machinery that allows us to get by in other domains of human endeavour? The question is bound up with many other questions of no less importance, questions concerning the uniqueness of the human mind, the course of biological evolution and the power of human culture. What is perhaps a little unusual about this question, however—unusual for a question whose answer concerns both those working in the sciences and the humanities—is that it can be phrased as a polar interrogative, i.e. as a question which admits of a yes or no response. And indeed the question has divided psychologists, linguists and the cognitive science community generally for many decades now, more or less into two camps. I would like to sketch the beginnings of an answer to this question—and others like it—in a way that does not pretend it can receive a simple yes or no response.

First of all, let me stress again that neural reuse is as well verified a phenomenon as one can expect in the cognitive sciences, and that it has left virtually no domain of psychology untouched. Neural reuse suggests that there is nothing so specialized in the cortex that it cannot be repurposed to meet new challenges while retaining its capacity for meeting old ones. In that regard, to be sure, what I am proposing is unapologetically on the side of those who maintain that language, as well as many other psychological capacities, are

not cognitively special—e.g. that there is no domain-specific "language organ" (cf. Chomsky 1980,39, 44; 1988, 159; 2002, 84-86).

And yet I would like to carefully distinguish this claim from the claim that there are no areas of the brain that subserve exclusively linguistic functions. The neuropsychological literature offers striking examples of what appear to be fairly clean dissociations between linguistic and nonlinguistic capacities, i.e. cases in which language processing capacities appear to be disrupted without impeding other cognitive abilities, and cases in which the reverse situation holds (Fedorenko et al. 2011; Hickok and Poeppel 2000; Poeppel 2001; Varley et al. 2005; Luria et al. 1965; Peretz and Coltheart 2003; Apperly et al. 2006). An example would be where the ability to hear words is disrupted, but the ability to recognize non-word sounds is spared (Hickok and Poeppel 2000; Poeppel 2001). Discussing such cases, Pinker and Jackendoff (2005, 207) add that "[c]ases of amusia and auditory agnosia, in which patients can understand speech yet fail to appreciate music or recognize environmental sounds…show that speech and non-speech perception in fact doubly dissociate." Although dissociations are to some extent compatible with reuse—indeed there is work suggesting that focal lesions can produce specific cognitive impairments within a range of nonclassical architectures (Plaut 1995)—and it is equally true that often the dissociations reported are noisy (Cowie 2008), still their very ubiquity needs to be taken seriously and accounted for in a more systematic fashion than many defenders of reuse have been willing to do (see e.g. Anderson 2010, 248; 2014, 46-48). After all, a good deal of support for

theories of reuse comes from the neuroimaging literature, which is somewhat ambiguous taken by itself. As Fedorenko et al. (2011, 16428) explain:

> standard functional MRI group analysis methods can be deceptive: two different mental functions that activate neighbouring but non-overlapping cortical regions in every subject individually can produce overlapping activations in a group analysis, because the precise locations of these regions vary across subjects, smearing the group activations. Definitively addressing the question of neural overlap between linguistic and nonlinguistic functions requires examining overlap within individual subjects, a data analysis strategy that has almost never been applied in neuroimaging investigations of high-level linguistic processing.

When Fedorenko and her colleagues applied this strategy themselves, they found that "most of the key cortical regions engaged in high-level linguistic processing are not engaged by mental arithmetic, general working memory, cognitive control or musical processing," and they think that this indicates "a high degree of functional specificity in the brain regions that support language" (2011, 16431). While I do not believe that claims of this strength have the least warrant—as I shall explain, functional specificity cannot be established merely by demonstrating that a region is selectively engaged by a task—these results do at least substantiate the dissociation literature in an interesting way and make it more difficult for

those who would prefer to dismiss the dissociations with a ready-made list of alternative explanations. Similar results were found by Fedorenko et al. (2012).

3. How Might Redundancy Feature In a Solution?

With rare exceptions (e.g. Friston and Price 2003; Barrett and Kurzban 2006; Jungé and Dennett 2010), redundancy has passed almost unnoticed in the philosophical and cognitive science literature. This is in stark contrast to the epigenetics literature, where redundancy and the related concept of degeneracy[5] have been explored to some depth (e.g. see Edelman and Gally 2001; Mason 2010; Whiteacre 2010; Deacon 2010; Iriki and Taoka 2012; Maleszka et al. 2013). The idea behind neural redundancy is that, for good evolutionary reasons (see below), the brain incorporates a large measure of redundancy of function. Brain regions (such as cortical columns and similar structures) fall in an iterative, repetitive and almost lattice-like arrangement in the cortex. Neighbouring columns have similar response properties: laminar and columnar changes are for the most part smooth—not abrupt—as one moves across the cortex, and adjacent modules do not differ markedly from one another in their basic structure and computations (if they really differ at all when taken in such

[5] Redundancy occurs when items have the same structure and function (i.e. are both isomorphic and isofunctional). Degeneracy occurs when items having *different* structures can perform the same function (i.e. are heteromorphic but isofunctional). Degeneracy implies genuine multiple realization (see Zerilli 2017b).

proximity). Regional *solitarity* is therefore not likely to be a characteristic of the brain (Anderson 2014, 141).[6] That is to say, we do not possess just one module for X, and one module for Y, but in effect several *copies* of the module for X, and several copies of the module for Y, all densely stuffed into the same cortical zones. As Buxhoeveden and Casanova (2002, 943) explain of neurons generally:

> In the cortex, more cells do the job that fewer do in other regions....As brain evolution paralleled the increase in cell number, a reduction occurred in the sovereignty of individual neurones; fewer of them occupy critical positions. As a consequence, plasticity and redundancy have increased. In nervous systems containing only a few hundred thousand neurones, each cell plays a more essential role in the function of the organism than systems containing billions of neurones.

The same principle very likely holds for functionally distinct groupings of neurons (i.e. cortical columns and like structures), as Jungé and Dennett (2010, 278) conjecture:

> It is possible that specialized brain areas contain a large amount of structural/computational redundancy (i.e., many neurons or collections of neurons

[6] The term "solitarity" is Anderson's, but while he concedes that solitarity will be "relatively rare," he does not appear to believe that anything particularly significant follows from this. See also Anderson (2010, 296).

> that can potentially perform the same class of functions). Rather than a single neuron or small neural tract playing roles in many high-level processes, it is possible that distinct subsets of neurons within a specialized area have similar competencies, and hence are redundant, but as a result are available to be assigned individually to specific uses....In a coarse enough grain, this neural model would look exactly like multi-use (or reuse).

This is plausibly why capacities which are functionally very closely related, but which for whatever reason are forced to recruit different neural circuits, will often be localized in broadly the same regions of the brain. For instance, first and second languages acquired early in ontogeny settle down in nearly the same region of Broca's area; and even when the second language is acquired in adulthood the second language is represented nearby within Broca's area (while artificial languages are not) (Kandel & Hudspeth 2013). The neural coactivation graphs of such composite networks must look very similar. Indeed these results suggest—and a redundancy model would predict—that two very similar tasks which are forced to recruit different neural circuits should exhibit similar patterns of activation. And this is more or less what we find (see below).

One might be tempted to think that redundancy and reuse pull in opposite directions. This is because whereas reuse posits that neural circuits get reused across different tasks and task categories, redundancy accommodates the likelihood of diverse

cognitive functions being activated by structurally and computationally equivalent circuits running in parallel: instead of a single circuit being reused across domains, two, three or more *copies* of that same circuit may be recruited differentially across those domains, such that no *single* circuit gets literally "re-used." But there is no substantive tension here. The redundancy account in truth *supplements* the reuse picture in a way that is consistent with the neuroimaging data, faithful to the core principle of reuse, and compatible with the apparent modularization and separate modifiability of technical and acquired skills in ontogeny. Evidence of the reuse of neural circuits to accomplish different tasks has, in fact, been adduced in aid of a theory which posits the reuse of the same neural *tokens* to accomplish these different tasks. Redundancy means we must accept that at least some of the time what we may actually be witnessing is reuse of the same *types* to accomplish these tasks. This does not diminish the standing of reuse. Let me explain.[7]

To the extent that a particular composite reuses types, and is dissociable pro tanto—residing in segregated brain tissue that is not active outside the domain in question—it is true that to that extent its constituents will *appear* to be domain-specific. But in this case looks will be deceiving. The classical understanding of domain specificity in effect *assumes* solitarity—that a module for X does something which no other module can do as well, or

[7] For a developmental twist on the type/token distinction invoked in the context of modular theorizing about the mind, see Barrett (2006).

that even *if* another module can do X as well, taken together these X-ing modules do not perform outside the X-domain. Here is an example of the latter idea (Bergeron 2007, 176):

> a pocket calculator could have four different division modules, one for dividing numbers smaller than or equal to 99 by numbers smaller than or equal to 99, a second one for dividing numbers smaller than or equal to 99 by numbers greater than 99, a third one for dividing numbers greater than 99 by numbers greater than 99, and a fourth one for dividing numbers greater than 99 by numbers smaller than or equal to 99. In such a calculator, these four capacities could all depend on (four versions of) the same algorithm. Yet, random damage to one or more of these modules in a number of such calculators could lead to observable (double) dissociations between any two of these functions.

Here, each module performs fundamentally the same algorithm, but in distinct hardware, such that dissociations are observable between any two functions. Notice, however, that none of these modules performs outside the "division" domain. This is what allows such duplicate modules to be considered domain-specific—they perform functions which, for all that they might run in parallel on duplicate hardware, are unique to a specific domain of operation, in this case division. If such modules could do work outside the division domain, they would lose the status of domain specificity, and acquire the status of domain neutrality (i.e. they would be domain-general). This is why a module that appears dedicated to a

particular function may not be domain-specific in the classical sense. Dedication is not the same as domain specificity, and redundancy, whether of calculator algorithms or neural circuits, explains why. A composite of neural regions will be dedicated without being domain-specific if its functional resources are accessible to other domains through the deployment (reuse) of neural surrogates (i.e. redundant or "proxy" tokens). In this case its constituents will be multi-potential but single-use (Jungé & Dennett 2010, 278), and the domain specificity on display somewhat cosmetic. To take an example with more immediate relevance to the brain, a set of cortical columns that are structurally and computationally similar may be equally suited for face recognition tasks, abstract-object recognition tasks, the recognition of moving objects, and so on. One of these columns could be reserved for faces, another for abstract objects, another for moving objects, and so on. What is noteworthy is that while the functional activation may be indistinguishable in each case, and the same *type* of resource will be employed on each occasion, a different *token* module will be at work at any one time. To quote Jungé and Dennett (2010, 278) again:

> In an adult brain, a given neuron [or set of neurons] would be aligned with only a single high-level function, whereas each area of neurons would be aligned with very many different functions.

Such modules (and composites) are for all intents and purposes *qualitatively* identical, though clearly not *numerically* identical, meaning that while they share their properties, they

are not *one and the same* (Parfit 1984). The evidence of reuse is virtually all one way when it comes to the pervasiveness of functional inheritance across cognitive domains. It may be that this inheritance owes to reuse of the same tokens (literal reuse) or to reuse of the same types (reuse by proxy), but the inheritance itself has been amply attested. This broader notion of reuse still offers a crucial insight into the operations of cognition, and I dare say represents a large part of the appeal of the original massive redeployment hypothesis (Anderson 2007c).

It is interesting to note in this respect that although detractors have frequently pointed out the ambiguity of neuroimaging evidence on account of its allegedly coarse spatial resolution (e.g. Carruthers 2010), suggesting that the same area will be active across separate tasks and task categories even if distinct but spatially adjacent and/or interdigitated circuits are involved in each case, this complaint can have no bearing on reuse by proxy. Fedorenko et al. (2011, 16431) take their neuroimaging evidence to support "a high degree of functional specificity in the brain regions that support language," but their results do not license this extreme claim. The regions they found to have been selectively engaged by linguistic tasks were all adjacent to the regions engaged in nonlinguistic tasks. Elementary considerations suggest that they have discovered a case of reuse by proxy involving language: the domains tested (mental arithmetic, general working memory, cognitive control and musical processing) make use of many of the same computations as high-level linguistic processing, even though they run them on duplicate hardware. Redundancy makes it is easy to see how fairly sharp dissociations could arise—knocking out one token module need

disrupt only one high-level operation: other high-level operations that draw on the same *type* of resource may well be spared.

The consequences of this distinction between literal reuse and reuse by proxy for much speculation about the localization and specialization of function are potentially profound. In cognitive neuropsychology the discovery that a focal lesion selectively impairs a particular cognitive function is routinely taken as evidence of its functional specificity (Coltheart 2011; Sternberg 2011). Even cognitive scientists who take a developmental approach to modularity, i.e. who concede that parts of the mind may be modular but stress that modularization is a developmental process, concede too much when they imply, as they frequently do, that modularization results in domain-specific modules (Karmiloff-Smith 1992; Prinz 2006; Barrett 2006; Cowie 2008; Guida et al. 2016). This is true in some sense, but not in anything like the standard sense, for redundancy envisages that developmental modules form a special class of neural networks, namely those which are *qualitatively* identical but *numerically* distinct. The appearance of modularization in development is thus fully compatible with deep domain interpenetration. In any event redundancy does not predict that all acquired skills will be modular. The evidence suggests that while some complex skills reside in at least partly dissociable circuitry, most complex skills are implemented in more typical neural networks, i.e. those consisting of literally shared parts.[8]

[8] This seems to be true regardless of whether the complex skills are innate or acquired.

4. What Else Might Redundancy Explain?

It is generally a good design feature of any system to have spare capacity. For instance, in engineered systems, "redundant parts can substitute for others that malfunction or fail, or augment output when demand for a particular output increases" (Whiteacre 2010, 14). The positive connection between robustness and redundancy in biological systems is also clear (Edelman and Gally 2001; Mason 2010; Whiteacre 2010; Iriki & Taoka 2012). So there are good reasons for evolution to have seen to it that our brains have spare capacity. But in the case of the brain and the cortex most especially, there are other reasons why redundancy would be an important design feature. It offers a solution to what Jungé and Dennett (2010, 278) called the "time-sharing" problem. It may also offer a solution to what I call the "encapsulation" problem.

The time-sharing problem arises when multiple simultaneous demands are made on the same cognitive resource. This is probably a regular occurrence. Here are just a few examples.

- Driving a car and holding a conversation at the same time: if it is true that some of the selfsame motor operations underlying aspects of speech production and comprehension are also required for the execution of sequenced or complex motor functions (Pulvermüller and Fadiga 2010; Graziano et al. 2002; MacNeilage 1998; Glenberg et al.

2008; Glenberg and Kaschak 2002; Glenberg et al. 2007; Greenfield 1991), as perhaps exemplified by driving a manual vehicle or operating complex machinery (e.g. playing the organ), how do we manage to pull this off?

- By reflecting the recursive structure of thought (Christiansen and Chater 2016, 51), the language circuits may redeploy a recursive operation simultaneously during sentence production. This might be the case during the formation of an embedded relative clause—the thought and its encoding may require parallel use of the same sequencing principle. Again, how do we manage this feat?
- If metarepresentational operations are involved in the internalization of conventional sound-meaning pairs, and also in the pragmatics and mindreading that carry on simultaneously during conversation, as argued by Suddendorf (2013), could this not simply be another instance of time-sharing? The example is contentious, but it still raises the question: how does our brain manage to do things like this?
- Christiansen and Chater's (2016) "Chunk and Pass" model of language processing envisages *multilevel* and *simultaneous* chunking procedures. As they put it, "the challenge of language acquisition is to learn a dazzling sequence of rapid processing operations" (2016, 116). What must the brain be like to allow for this dazzling display?

Explaining these phenomena is difficult. Indeed when dealing with clear (literal) instances of reuse, results from the interference paradigm show that processing bottlenecks are inevitable—true multi-tasking is impossible. Redundancy offers a natural explanation of how

the brain overcomes the time-sharing problem. It explains, in short, how we are able to walk and chew gum at the same time.

Redundancy might also offer a solution to what I have called the encapsulation problem. The neural networks that implement cognitive functions are not likely to be characterized by informational encapsulation if they share their nodes with networks implementing other cognitive functions. This is because in sharing their nodes with these other systems they will prima facie have access to the information stored and manipulated by those other systems (Anderson 2010, 300). If, then, overlapping brain networks must share information (Pessoa 2016, 23), it would be reasonable to suppose that central and peripheral systems do *not* overlap. For peripheral systems, which are paradigmatically fast and automatic, would not be able to process inputs as efficiently if there were a serious risk of central system override—i.e. of beliefs and other central information getting in the way of automatic processing. But we know from the neuroimaging literature that quite often the brain networks implementing central and peripheral functions *do* overlap. This is puzzling in light of the degree of cognitive impenetrability that certain sensory systems still seem to exhibit—limited though it may be. If it is plausible to suppose that the phenomenon calls for segregated circuitry, redundancy could feature in a solution to the puzzle, since it naturally explains how the brain can make parallel use of the same resources. Neuroimaging maps might well display what appear to be overlapping brain regions between two tasks (one involving central information, the other involving classically peripheral operations), but the

overlap would not exist—there would be distinct albeit adjacent or interdigitated and nearly identical circuits recruited in each case. Of course there may be other ways around the encapsulation problem that do not require segregated circuitry: the nature and extent of the overlap is presumably important. But clearly redundancy opens up some fascinating explanatory possibilities.

To the extent that acquired skills must overcome both the time-sharing problem as well as the encapsulation problem—for acquired competencies are often able to run autonomously of central processes—we might expect that their neural implementations incorporate redundant tissue. In concluding, let me illustrate this point by offering a gloss on a particular account of how skills and expertise are acquired during development elaborated by Guida et al. (2016) and Anderson (2014). The process involved is called "search" (Anderson 2014). Search is an exploratory synaptogenetic process, "the active testing of multiple neuronal combinations until finding the most appropriate one for a specific skill, i.e., the neural niche of that skill" (Guido et al. 2016, 13). The theory holds that in the early stages of skill acquisition, the brain must search for an appropriate mix of brain areas, and does so by recruiting relatively widely across the cortex. When expertise has finally developed, a much narrower and more specific network of brain areas has been settled upon, such that "[a]s a consequence of their extended practice, experts develop domain-specific knowledge structures" (Guido et al. 2016, 13). The gloss (and my hunch) is this: first, that repeated practice of a task that requires segregation (to get around time-

sharing and encapsulation issues) will in effect *force* search into redundant neural territory (Karmiloff-Smith 1992; Barrett 2006; Barret and Kurzban 2006); second, that search will recruit idle or relatively underutilized circuits in preference to busy ones as a general default strategy. Guido et al. (2016) cite evidence that experts' brains reuse areas for which novices' brains make only limited use: "Whereas novices use episodic long-term memory areas (e.g., the mediotemporal lobe) for performing long-term memory tasks, experts are able to (re)use these areas also for performing working-memory tasks" (Guido et al. 2016, 14). Guido and colleagues, in agreement with Anderson (2014), seem to have literal reuse in mind. But the same evidence they cite is consistent with reuse by proxy. As Barrett and Kurzban (2006, 639) suggest, echoing a similar suggestion by Karmiloff-Smith (1992), a developmental system

> could contain a procedure or mechanism that partitioned off certain tasks—shunting them into a dedicated developmental pathway—under certain conditions, for example, when the cue structure of repeated instances of the task clustered tightly together, and when it was encountered repeatedly, as when highly practiced....Under this scenario, reading could still be recruiting an evolved system for object recognition, and yet phenotypically there could be *distinct modules* for reading and for other types of object recognition.

5. Conclusion

It is true that language and other cognitive skills frequently dissociate from other skills, but redundancy puts this sort of modularization in its proper context. Redundancy predicates functional inheritance across tasks and task categories even when the tasks are implemented in spatially segregated neural networks. Thus dissociation evidence alone does not always indicate true functional specificity. In particular, these dissociations provide no evidence that language is cognitively special vis-à-vis other cognitive domains.

References

Anderson, Michael L. 2007a. "Evolution of Cognitive Function via Redeployment of Brain Areas." *The Neuroscientist* 13:13-21.

——2007b. "Massive Redeployment, Exaptation, and the Functional Integration of Cognitive Operations." *Synthese* 159 (3): 329-345.

——2007c. "The Massive Redeployment Hypothesis and the Functional Topography of the Brain." *Philosophical Psychology* 21 (2): 143-174.

——2008. "Circuit Sharing and the Implementation of Intelligent Systems." *Connection Science* 20 (4): 239-251.

——2010. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences* 33 (4): 245-266; discussion 266-313.

——2014. *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Anderson, Michael L., and Barbara L. Finlay. 2014. "Allocating Structure to Function: The Strong Links Between Neuroplasticity and Natural Selection." *Frontiers in Human Neuroscience* 7:1-16.

Apperly, I.A., D. Samson, N. Carroll, S. Hussain, and G. Humphreys. 2006. "Intact First- and Second-Order False Belief Reasoning in a Patient with Severely Impaired Grammar." *Social Neuroscience* 1 (3-4): 334-348.

Barrett, H. Clark. 2006. "Modularity and Design Reincarnation." In *The Innate Mind Volume 2: Culture and Cognition*, ed. Peter Carruthers, Stephen Laurence, and Stephen P. Stich, 199-217. New York: Oxford University Press.

Barrett, H. Clark, and Robert Kurzban. 2006. "Modularity in Cognition: Framing the Debate." *Psychological Review* 113 (3): 628-647.

Bergeron, Vincent. 2007. "Anatomical and Functional Modularity in Cognitive Science: Shifting the Focus." *Philosophical Psychology* 20 (2): 175-195.

Buxhoeveden, Daniel P., and Manuel F. Casanova. 2002. "The Minicolumn Hypothesis in Neuroscience." *Brain* 125:935-951.

Carruthers, Peter. 2006. The *Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Casasanto, D., and K. Dijkstra. 2010. "Motor Action and Emotional Memory." *Cognition* 115 (1): 179-185.

Chomsky, Noam. 1980. *Rules and Representations*. New York: Columbia University Press.

——1988. *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.

——2002. *On Nature and Language*. New York: Cambridge University Press.

Christiansen, Morten H., and Nick Chater. 2016. *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge, MA: MIT Press.

Coltheart, Max. 1999. "Modularity and Cognition." *Trends in Cognitive Sciences* 3 (3): 115-120.

——2011. "Methods for Modular Modelling: Additive Factors and Cognitive Neuropsychology." *Cognitive Neuropsychology* 28 (3-4): 224-240.

Cosmides, Leda, and John Tooby. 1994. "Origins of Domain Specificity: The Evolution of Functional Organization." In *Mapping the World: Domain Specificity in Cognition and Culture*, ed. L. Hirschfield, and S. Gelman, 85-116. New York: Cambridge University Press.

Cowie, Fiona. 2008. "Innateness and Language." In *The Stanford Encyclopedia of Philosophy*, winter 2016, ed. E.N. Zalta. <http://plato.stanford.edu/archives/win2016/entries/innateness-language/>

de Almeida, Roberto G., and Lila R. Gleitman, eds. 2018. *On Concepts, Modules, and Language: Cognitive Science at its Core*. New York: Oxford University Press.

Deacon, Terrence W. 2010. "A Role for Relaxed Selection in the Evolution of the Language Capacity." *Proceedings of the National Academy of Sciences of the United States of America* 107: 9000-9006.

Dehaene, Stanislas. 2005. "Evolution of Human Cortical Circuits for Reading and Arithmetic: The 'Neuronal Recycling' Hypothesis." In *From Monkey Brain to Human Brain*, eds. Stanislas Dehaene, J.R. Duhamel, M.D. Hauser, and G. Rizzolatti, 133-157. Cambridge, MA: MIT Press.

Edelman, Gerald M., and Joseph A. Gally. 2001. "Degeneracy and Complexity in Biological Systems." *Proceedings of the National Academy of Sciences of the United States of America* 98 (24): 13763-13768.

Eliasmith, Chris. 2015. "Building a Behaving Brain." In *The Future of the Brain*, ed. Gary Marcus, and Jeremy Freeman, 125-136. Princeton: Princeton University Press.

Fedorenko, Evelina, Michael K. Behr, and Nancy Kanwisher. 2011. "Functional Specificity for High-Level Linguistic Processing in the Human Brain." *Proceedings of the National Academy of Sciences of the United States of America* 108 (39): 16428-16433.

Fedorenko, Evelina, John Duncan, and Nancy Kanwisher. 2012. "Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area." *Current Biology* 22 (21): 2059-2062.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Friston, Karl J., and Cathy J. Price. 2003. "Degeneracy and Redundancy in Cognitive Anatomy." *Trends in Cognitive Sciences* 7 (4): 151-152.

Gauthier, I., T. Curran, K.M. Curby, and D. Collins. 2003. "Perceptual Interference Supports a Non-Modular Account of Face Processing." *Nature Neuroscience* 6 (4): 428-432.

Glenberg, A.M., M. Brown, and J.R. Levin. 2007. "Enhancing Comprehension in Small Reading Groups Using a Manipulation Strategy." *Contemporary Educational Psychology* 32:389-399.

Glenberg, A.M., and M.P. Kaschak. 2002. "Grounding Language in Action." *Psychonomic Bulletin and Review* 9:558-565.

Glenberg, A.M., M. Sato, and L. Cattaneo. 2008. "Use-Induced Motor Plasticity Affects the Processing of Abstract and Concrete Language." *Current Biology* 18 (7): R290-291.

Godfrey-Smith, Peter. 2001. "Three Kinds of Adaptationism." In *Adaptationism and Optimality*, ed. Steven H. Orzack, and Elliott Sober, 335-357. Cambridge: Cambridge University Press.

Graziano, M.S.A., C.S.R. Taylor, T. Moore, and D.F. Cooke. 2002. "The Cortical Control of Movement Revisited." *Neuron* 36:349-362.

Greenfield, P.M. 1991. "Language, Tools and Brain: The Ontogeny and Phylogeny of Hierarchically Organized Sequential Behavior." *Behavioral and Brain Sciences* 14 (4): 531- 551; discussion 551-595.

Guida, Alessandro, Guillermo Campitelli, and Fernand Gobet. 2016. "Becoming an Expert: Ontogeny of Expertise as an Example of Neural Reuse." *Behavioral and Brain Sciences* 39:13-15.

Hickok, G., and David Poeppel. 2000. "Towards a functional neuroanatomy of speech perception." *Trends in Cognitive Sciences* 4 (4): 131-138.

Iriki, Atsushi, and Miki Taoka. 2012. "Triadic (ecological, neural, cognitive) niche construction: A scenario of human brain evolution extrapolating tool use and language from the control of reaching actions." *Philosophical Transactions of the Royal Society B* 367: 10-23.

Jungé, Justin A., and Daniel C. Dennett. 2010. "Multi-Use and Constraints from Original Use." *Behavioral and Brain Sciences* 33 (4): 277-278.

Kandel, E.R., and A.J. Hudspeth. 2013. "The Brain and Behavior." In *Principles of Neural Science*, ed. E.R. Kandel, J.H. Schwartz, T.M. Jessell, S.A. Siegelbaum, and A.J. Hudspeth, 5-20. New York: McGraw-Hill.

Karmiloff-Smith, Annette. 1992. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Leo, Andrea, Giulio Bernardi, Giacomo Handjaras, Daniela Bonino, Emiliano Ricciardi, and Pietro Pietrini. 2012. "Increased BOLD Variability in the Parietal Cortex and Enhanced Parieto-Occipital Connectivity During Tactile Perception in Congenitally Blind Individuals." *Neural Plasticity* 2012:1-8 doi: 10.1155/2012/720278.

Luria, A.R., L.S. Tsvetkova, and D.S. Futer. 1965. "Aphasia in a Composer (V.G. Shebalin)." *Journal of the Neurological Sciences* 2 (3): 288-292.

MacNeilage, P.F. 1998. "The Frame/Content Theory of Evolution of Speech Production." *Behavioral and Brain Sciences* 21 (4): 499-511; discussion 511-546.

Maleszka, Ryszard, Paul H. Mason, and Andrew B. Barron. 2013. "Epigenomics and the Concept of Degeneracy in Biological Systems." *Briefings in Functional Genomics* 13 (3): 191-202.

Mason, Paul H. 2010. "Degeneracy at Multiple Levels of Complexity." *Biological Theory* 5 (3): 277-288.

McGeer, Victoria. 2007. "Why Neuroscience Matters to Cognitive Neuropsychology." *Synthese* 159:347-371.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Pessoa, Luiz. 2016. "Beyond Disjoint Brain Networks: Overlapping Networks for Cognition and Emotion." *Behavioral and Brain Sciences* 39:22-24.

Peretz, Isabelle., and Max Coltheart. 2003. "Modularity of music processing." *Nature Neuroscience* 6:688-691.

Pinker, Steven, and Ray Jackendoff. 2005. "The Faculty of Language: What's Special About It?" *Cognition* 95:201-236.

Plaut, David C. 1995. "Double Dissociation Without Modularity: Evidence from Connectionist Neuropsychology." *Journal of Clinical and Experimental Psychology* 17 (2): 291-321.

Poeppel, David. 2001. "Pure Word Deafness and the Bilateral Processing of the Speech Code." *Cognitive Science* 21 (5): 679-693.

Prinz, Jesse J. 2006. "Is the Mind Really Modular?" In *Contemporary Debates in Cognitive Science*, ed. R. Stainton, 22-36. Oxford: Blackwell.

Pulvermüller, Friedmann, and Luciano Fadiga. 2010. "Active Perception: Sensorimotor Circuits as a Cortical Basis for Language." N*ature Reviews Neuroscience* 11:351-360.

Sternberg, Saul. 2011. "Modular Processes in Mind and Brain." *Cognitive Neuropsychology* 28 (3-4): 156-208.

Striem-Amit, Ella, and Amir Amedi. 2014. "Visual Cortex Extrastriate Body-Selective Area Activation in Congenitally Blind People 'Seeing' by Using Sounds." *Current Biology* 24:1-6.

Suddendorf, Thomas. 2013. *The Gap: The Science of What Separates Us from the Animals*. New York: Basic Books.

Varley, R.A., N.J.C. Klessinger, C.A.J. Romanowski, and M. Siegal. 2005. "Agrammatic But Numerate." *Proceedings of the National Academy of Sciences of the United States of America* 102:3519-3524.

Whiteacre, James M. 2010. "Degeneracy: A Link Between Evolvability, Robustness and Complexity in Biological Systems." *Theoretical Biology and Medical Modelling* 7 (6): 1-17.

Zerilli, John. 2017a. "Against the 'System' Module." *Philosophical Psychology* 30 (3): 235-250.

——2017b. "Multiple Realization and the Commensurability of Taxonomies." *Synthese* (https://doi.org/10.1007/s11229-017-1599-1).

Explanatory Conditionals

Holger Andreas

University of British Columbia

Accepted for publication in the December 2019 proceedings issue of
Philosophy of Science

Abstract

The present paper aims to complement causal model approaches to causal explanation by Woodward [15], Halpern and Pearl [5], and Strevens [14]. It centres on a strengthened Ramsey Test of conditionals: $\alpha \gg \gamma$ iff, after suspending judgment about α and γ, an agent can infer γ from the supposition of α (in the context of further beliefs in the background). It has been shown by Andreas and Günther [1] that such a conditional can be used as starting point of an analysis of 'because' in natural language. In what follows, we shall refine this analysis so as to yield a fully fledged account of (deterministic) causal explanation.

1 Introduction

The present paper aims to complement causal model approaches to causal explanation by Woodward [15], Halpern and Pearl [5], and Strevens [14]. It does so by carrying on a conditional analysis of the word 'because' in natural language by Andreas and Günther [1]. This analysis centres on a strengthened Ramsey Test of conditionals:

> $\alpha \gg \gamma$ iff, after suspending judgment about α and γ, an agent can infer γ from the supposition of α (in the context of further beliefs in the background).

Using this conditional, we can give a logical analysis of because:

$$\textit{Because } \alpha, \gamma \text{ (relative to } K) \quad \text{iff} \quad \alpha \gg \gamma \in K \quad \text{and} \quad \alpha, \gamma \in K$$

where K designates the belief set of the agent. In what follows, we shall refine this analysis by further conditions so as to yield a fully fledged analysis of deterministic causal explanations. The logical foundations of the belief changes that define the conditional $\gg$ are explicated using AGM-style belief revision theory [3].

Why do we think that causal model approaches to causal explanation are incomplete? Halpern and Pearl [5] and Woodward [15] are the most prominent elaborations of a causal model approach in contemporary philosophy of science. Halpern and Pearl [4] have devised a precise semantics of causal models that centres on structural equations. Such an equation represents causal dependencies between variables in a causal model:

$$X := \phi(Y_1, \ldots, Y_n).$$

In this schema of a structural equation, the variable X causally depends on the variables $Y_1, \ldots, Y_n$. It proved highly useful to represent the causal dependencies of a causal model by a causal graph. Woodward's account of causal explanations in [15] heavily relies on such graphs.

In the definition of causation by Halpern and Pearl [4], there is no explanation of what it is for a variable to causally depend *directly* on certain other variables. This approach merely defines complex causal relations in terms of elementary causal dependencies, just as truth-conditional semantics defines the semantic values of complex sentences in terms of a truth-value assignment to the atomic formulas. And the corresponding account of causal explanation in Halpern and Pearl [5] inherits the reliance on elementary causal dependencies (which are assumed to be antecedently given) from the analysis of causation.

Woodward [15] explains the notion of a direct cause in terms of interventions, but the notion of an intervention is always relative to a causal graph so that some knowledge about elementary causal dependencies must be antecedently. An account of causal explanation in terms of elementary causal dependencies is certainly valuable and insightful, but it is not the full story. At least, we should not give up on a more comprehensive account of causation and causal explanation too quickly (cf. Spohn [11] and Paul and Hall [8]).

The kairetic account of explanation by Strevens [14, 13] makes essential use of causal models as well, but works with a more liberal notion of such a model. In

this account, a set of propositions *entail* an explanandum E in a causal model only if this entailment corresponds to a "real causal process by which E is causally produced" [13, p. 165]. Causal models are assumed to be founded in physical facts about causal influence, and it is assumed that these facts "can be read off the true theory of everything" [13, p. 165].

The kairetic account is conceptually incomplete in a manner akin to the approaches by Halpern and Pearl [5] and Woodward [15]. This account account leaves open how we can discriminate between causal relations of logical entailment and non-causal ones in a true theory about the world, be it complete or incomplete. Relatedly, the account comes with a number of placeholders that remain unspecified. For example, causal models are assumed to define a relation of logical entailment that represents actual causal relations. What is the language of such a causal model? How is the mathematics underlying some causal processes represented? What are the distinctive properties of causal relations of logical entailment? In what follows, we shall make an attempt at answering these questions, focusing in particular on a characterization of logical entailment with a causal meaning. For this characterization, we define an explanatory conditional $\gg$, but impose also non-logical conditions on the explanans and the explanandum.

Our final account of logical entailment with a causal meaning takes as input a non-modal, first-order representation of our theories and beliefs about the world. Modality and explanatory conditionals come into play through a refined and strengthened variant of the Ramsey Test. The final analysis is thus relative to an epistemic state that represents certain theories and beliefs about the world. This yields a smooth epistemology of causal explanations insofar as we seem to know what beliefs and theories we have about the world. The analysis is thus very much in the spirit of related accounts of explanation and causation by Gärdenfors [3, Ch. 8-9] and Spohn [11]. Note, however, that we are not committed to the view that causation and explanation are epistemic concepts. If we are given an epistemic state that contains the true theory of everything, our analysis explains how we can read the causal structure off such a theory or a portion thereof.

2 Belief Changes and the Ramsey Test

2.1 Belief Changes: Basic Ideas

AGM-style belief revision theory provides us with a precise semantics of belief changes for the Ramsey Test. Let us therefore make ourselves familiar with the basic ideas of this theory. In the AGM framework, one distinguishes three types of belief change of a belief set K by a formula α:

(1) Expansions $K + \alpha$

(2) Revisions $K * \alpha$

(3) Contractions $K \div \alpha$.

An expansion of K by α consists in the addition of a new belief α to the belief system K. This operation is not constrained by any considerations as two whether the new epistemic input α is consistent with the the set K of present beliefs. Hence, none of the present beliefs is retracted by an expansion. $K + \alpha$ designates the expanded belief set.

A revision of K by α, by contrast, can be described as the *consistent integration* of a new epistemic input α into a belief system K. If α is consistent with K, it holds that $K + \alpha = K * \alpha$, i.e., the revision by α is equivalent with the expansion by α. If, however, α is not consistent with K, some of the present beliefs are to be retracted, as a consequence of adopting the new epistemic input. $K * \alpha$ designates the revised system of beliefs.

A contraction of K by α, finally, consists in retracting a certain formula α from the presently accepted system of beliefs. This operations will be used to define the *suspension of judgement about* α in our strengthened version of the Ramsey Test. $K \div \alpha$ designates the belief set after the retraction of α.

In some contexts, it is helpful to distinguish between the belief system K and the epistemic state S that underlies it. Henceforth, we shall make this distinction, and write $K(S)$ for the belief system K of the epistemic state S.

Belief changes can be defined in various ways. A large number of different belief revision schemes have been developed in the spirit of the original AGM theory. We shall assume that epistemic states are represented by *belief bases*. In symbols, $S = H$. A belief base H is a set of formulas that represent the explicit beliefs of

an agent. Belief base revision schemes are guided by the idea that the inferential closure of a belief base H gives us the belief set K of H:

$$K(H) =_{df} Inf(H).$$

K contains all beliefs of the epistemic state H, i.e, the explicit beliefs and those beliefs that the agent is committed to accept because they are derivable from the explicit beliefs. Inf is an inferential closure operation that may or may not be given by the consequence operation of classical logic. Henceforth, we shall assume that $K(H) = Cn(H)$, i.e., the belief set $K(H)$ is the classical logical closure of H.

The definition of an expansion is straightforward for belief bases:

$$K(H) + \alpha =_{df} K(H + \alpha)$$

where $H + \alpha$ stands for adding the new epistemic input to the belief base H.

Note that we can define revisions in terms of contractions and expansions:

$$K(S) * \alpha = (K(S) \div \neg\alpha) + \alpha. \qquad \text{(Levi identity)}$$

Once we have retracted α, we obtain a belief set $K(S')$ that is consistent with α. Hence, we have $K(S')*\alpha = K(S')+\alpha$. Such are the basic ideas about belief changes that will be used in our strengthened Ramsey Test and the subsequent analysis of explanation.[1]

2.2 The Ramsey Test

The Ramsey Test is an epistemic approach to conditionals. Its core idea has been expressed most clearly by Richard Stalnaker [12, p. 102]:

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent), finally, consider whether or not the consequent is then true.

It was then Peter Gärdenfors [3] who translated this test into the language of belief changes and who insisted more forcefully than Stalnaker [12] on an epistemic understanding of conditionals. Using the AGM framework, he was able to explicitly define a semantics of conditionals in terms of belief changes:

$$\alpha > \gamma \in K(S) \text{ iff } \gamma \in K(S) * \alpha$$

[1] For further details, the reader is referred to Hansson [6].

where $>$ designates the conditional connective. Recall that $K(S) * \alpha$ designates the revision of the beliefs of an epistemic state S with the formula α. So the Ramsey Test defines that a conditional $\alpha > \gamma$ is to be accepted in a belief system $K(S)$ iff it is true of $K(S)$ that the consequent γ is in $K(S)$ when revised by the antecedent α.

2.3 Strengthening the Ramsey Test

Inspired by the work of Hans Rott [9] on the logical analysis of 'because', we define a conditional $\gg$ with the following intuitive meaning: $\alpha \gg \gamma$ iff, after suspending any beliefs in $K(S)$ as to whether α and γ are true or false, it holds that $\gamma \in K(S) * \alpha$. The evaluation of $\alpha \gg \gamma$, thus, consists of two steps: (i) contracting $K(S)$ in such a manner that we become indeterminate about α and γ; (ii) testing whether or not γ is in $K(S) * \alpha$. In more formal terms:

Definition 1. Belief function $B(\alpha)$
Let T be some arbitrary classical tautology and α a formula.

$$B(\alpha) = \begin{cases} \alpha & \text{if } \alpha \in K(S) \\ \neg\alpha & \text{if } \neg\alpha \in K(S) \\ \neg T & \text{otherwise.} \end{cases}$$

$$\alpha \gg \gamma \in K(S) \text{ iff } \alpha \in (K(S) \div B(\alpha) \vee B(\gamma)) * \alpha. \tag{SRT}$$

Contracting $K(S)$ by $B(\alpha) \vee B(\gamma)$ results in a belief system $K(S')$ that does neither contain α nor $\neg\alpha$, nor γ, nor $\neg\gamma$, provided that α and γ are contingent. It thus represents the operation of suspending judgement about the truth and falsity of α, γ. Hence, $\alpha \gg \gamma$ iff $\gamma \in K(S') * \alpha$.

Using $\neg\alpha \notin K(S) \div B(\alpha) \vee B(\gamma)$ and the Levi identity, we obtain:

$$\alpha \gg \gamma \text{ iff } (K(S) \div B(\alpha) \vee B(\gamma)), \alpha \vdash \gamma$$

where $\vdash$ stands for the relation of logical consequence in classical logic. $\alpha \gg \gamma$, thus, means that the consequent γ is inferable from the antecedent α, together with other explicit beliefs, after judgement has been suspended about γ and α.

2.4 Ramsey Test Explanations

Having strengthened the Ramsey Test, we are in a position to give a preliminary account of explanation:

Definition 2. Explanation (preliminary account)
Let S be an epistemic state that is represented by a belief base. The set A of antecedent conditions and the set G of generalizations *explain* the fact F - relative to S - iff

(E1) For all $\alpha \in A$, all $\gamma \in G$, and all $\beta \in F$: $\alpha, \gamma, \beta \in K(S)$.

(E2) For all non-empty $A' \subseteq A$, $\bigwedge A' \gg \bigwedge F \in K(S)$.

$\bigwedge C$ is shorthand for $\bigwedge_{\phi \in C} \phi$ and, hence, designates a conjunction of all members of a set C. We do not merely require that $\bigwedge A \gg \bigwedge F$ to ensure that all propositions of A are relevant for the explanandum. F is a set because the description of the explanandum may be complex.

This preliminary account bears substantial commonalities with Strevens [14] and accounts in the logical empiricist tradition. All of these accounts require that the explanandum be entailed by a set of antecedent conditions, together with a set of generalizations or laws. In line with Strevens [14], we are aiming to capture the intuition that the explanandum is *causally produced* by the antecedent conditions of the explanation. For this to be achieved, we need to impose further constraints on the Ramsey Test, the explanans, and explanandum.

3 Causation

3.1 Time

For C to be a cause of E, it must hold that

$$C \gg E \in K(S). \tag{C1}$$

This condition allows us to capture a large range of causal relations. It is too liberal, however. If a theory T is deterministic and time-symmetric, we are not only able to infer the effect from the cause, but also the other way around. Classical mechanics is such a theory.

Now, the idea of production seems to imply a temporal asymmetry between the producing event and the effect: the cause must precede its effect. Hence,

$$t(C) < t(E) \tag{C2}$$

where $t(C)$ is a function that yields the time at which the event C occurs. This condition expresses an old Humean dictum, which is also central to Spohn's [11] ranking-theoretic analysis. It helps not mistake effects for causes in deterministic, time-symmetric theories. If A, B, or A and B are temporally extended events, we take $t(A) < t(B)$ to mean that A comes into being before B, while A and B may well overlap. In next section, we shall see that condition (C2) is too restrictive and so needs to be liberalized – not only because of ideas about backward causation.

3.2 Levels of Theoreticity

The temporal asymmetry between cause and effect appears to hold consistently in everyday contexts. It does not always hold in scientific contexts. For example, we say that forces produce accelerations and that electromagnetic fields produce forces upon charged particles, without there being a temporal delay between, respectively, forces and accelerations, and electromagnetic fields and electromagnetic forces. Is this accidental? If not, what is the distinctive relation that allows us to recognize a causal direction in the latter cases?

We conjecture that an ordering of *theoreticity* plays an important role here. Forces are higher up in this ordering than accelerations because Newtonian mechanics determines the meaning of force and mass on the basis of the concepts of space and time, but not vice versa.[2] The notion of an ordering of theoreticity has been made precise recently by Schurz [10]. The core idea of this definition is that a concept C_1 is higher up in this ordering than a concept C_2 iff there is a chain of theories that define measurement procedures that lead from C_1 to empirical concepts via C_2. The ordering of theoreticity thus defined may well be partial. For lack of space, we cannot go further into the details.

Note that there is also a philosophical motivation for recognizing a connection between theoreticity and causation. A major motivation for introducing theoretical concepts is to give a unified account of certain phenomena. In light of unificationist ideas about explanation by Kitcher [7], this implies that theoretical concepts are – for conceptual reasons – supposed to help devise explanations. If we then further

[2]This is so on all formal semantics of theoretical concept.

assume a close connection between causation and explanation, we can infer that theoretical concepts are supposed to provide causal explanations.

Using our conjecture concerning theoreticity and causation, we are able to liberalize condition (C2) as follows:

$$t(C) < t(E) \text{ or } c(E) <_t c(C) \qquad (C2^*)$$

In this notation, $c(A)$ gives us the set of concepts in terms of which the sentence A is expressed. $c(E) <_t c(C)$ is intended to mean that there is at least one concept in $c(C)$ that is higher up in the theoreticity ordering than all the concepts of $c(E)$, while there is no concept in $c(C)$ that is below a concept in $c(E)$ in the sense of this ordering.

3.3 Joint Effects

Joint effects of a common cause can pose a problem for an inferential approach to causation and explanation. Take the following neuron diagram [8, p. 71]:

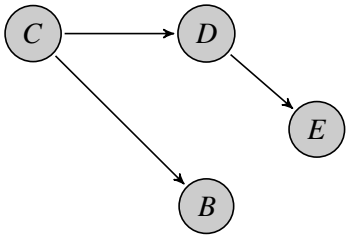


Figure 1

C fires and, thereby, sends signals to D and B so that D and B are excited. E is excited in the course of receiving a signal from D. Intuitively, the excitation of B is not a cause of the excitation of E. However, the excitation of B strongly conditionally implies – in the sense of (SRT) – the excitation of E.

Counterfactual approaches solve the problem by excluding *backtracking* counterfactuals [8, 71-72]. We can adopt a similar strategy. The counterintuitive result about joint effects is avoided if all inferences that lead from the presumed cause C to the putative effect E are non-backtracking. We can make this idea precise in proof-theoretic terms. Recall that any inferential step in a natural deduction proof

consists of a set P of premises and a conclusion C, where P may contain subproofs as premises. This in mind, we can define the notion of a *forward-directed proof*. Such a proof conforms to the temporal order of events in the following sense:

Definition 3. $H \vdash_F C$
Let H be a set of formulas and C be a formula. Only literals and conjunctions of literals are taken to assert the occurrence of an event. We say there is a *forward-directed natural deduction proof* of C from H – in symbols $H \vdash_F C$ – iff there is a natural deduction proof of C from H such that (i) for all inferential steps P/I (of the main proof and any subproof), if I asserts the occurrence of an event, then this event does not precede any event that is asserted by a premise in P or by a premise in a subproof that is a member of P, and (ii) the assumption of any subproof is consistent with H.

Using this notion of a forward-directed proof, we can impose a temporal constraint on our Ramsey Test:

$$A \gg_F C \in K(S) \text{ iff } (K(S) \div B(A) \vee B(C)), A \vdash_F C. \qquad (SRT_F)$$

That is, C is a forward-directed strong conditional implication of A iff there is a forward-directed proof of C from A and the explicit beliefs of the epistemic state S, after suspending judgement on the antecedent A and the consequent C. Now, we require there to be a forward-directed proof of the putative effect from the presumed cause:

$$C \gg_F E \in K(S). \qquad (C1^*)$$

This condition solves the problem of joint effects. For, the inference from the firing of B to the firing of C is backward-directed. Hence, there is no forward-directed proof of E from A. Note that a forward-directed proof merely excludes backward-directed inferences. Such a proof may still involve inferences where the events asserted in the premises and the conclusion are simultaneous.

3.4 Spurious Causation

Unfortunately, there is another problem with the causal scenario of Figure 1. As there is a stable correlation between the firing of B and E, it is reasonable to have the implication $B \rightarrow E$ as a generalization in the belief base. Hence, $B \gg E$. If the firing of B precedes that of E, the inference is forward-directed, and so $B \gg_F E$

holds as well. So we would have to consider the firing of B as a cause of the firing of E, which is counterintuitive. This is the problem of spurious causation.

The paradigm of a spurious cause is the drop of the barometer which does not count as a genuine cause of any storm. Since, however, the correlation between the drop of the barometer and the storm is commonly considered to be probabilistic, this example is not well suited for the present analysis of deterministic causation. Here is a better example: a causal analysis of a thunderstorm should distinguish at least three events: (i) the electrical discharge between a cloud and the ground, (ii) the flash of the lightning, and (iii) the thunder. Physics tells us that the electrical discharge between a cloud and the ground is – via the rapid production of heat within the region of the air where electricity is conducted – the common cause of the flash and the thunder. Physics would deny that the flash is a genuine cause of the thunder. So we have to conclude that the flash is a mere spurious cause of the thunder. But the event of the flash satisfies conditions ($C1^*$) and ($C2^*$) with respect to the thunder. So we must further refine our analysis.

Why is the electrical discharge rather than the flash a genuine cause of the thunder that temporally follows? Let us take a closer look at the various inferential paths in this causal scenario. The flash and the electrical discharge of the lightning are alike in that we can infer from their occurrence the occurrence of the thunder. The inferential paths, however, differ from one another. We can infer the thunder from the electrical discharge, in a forward-directed manner, using only generalizations that are fundamental, or non-redundant, *laws of nature*. In essence, it is the laws of electrodynamics, atomic theory, and acoustics that are used in this path. By contrast, there is no forward-directed inferential path from the flash to the thunder such that all generalizations thereof are fundamental laws of nature. The inference from the flash of the lightning to the thunder is either not forward-directed (when it goes via the electrical discharge) or not based on fundamental laws of nature (when the thunder is directly inferred from the flash, without going through the common cause of the flash and the thunder). Figure 2 may help distinguish the two inferential paths:

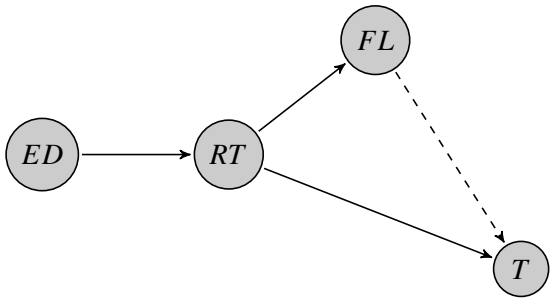


Figure 2

The symbols have their obvious meanings: ED stands for *electrical discharge*, RT for *rise of temperature*, FL for flash of lightning, and T for thunder.

Let us be more precise about the notion of a fundamental, or non-redundant, law of nature. To make our approach as parsimonious as possible and to avoid conflict with alternative accounts of laws of nature, we prefer the notion of a *non-redundant generalization* over the notion of a law of nature. Yet, we explicate the former notion in terms of a *best system account* of laws of nature. In fact, a very weak understanding of what a best system is suffices to make the crucial distinction:

Explanation 1. Non-redundant generalization
A generalization is non-redundant iff it is a member of a deductive system that does a good job at balancing strength and simplicity and that is not clearly inferior to an alternative system.

This formulation acknowledges that the optimization of strength and simplicity cannot be accomplished in a straightforward manner.[3] However, it is strong enough to see that the generalization asserting a regular connection between a flash at the sky and a thunder is *redundant*. For, (i) it is uncontroversial that the laws of electrodynamics, atomic theory, acoustics, etc. (i. e. the laws that are used in the inferential path from the flash to the thunder via the electrical discharge) are members of any deductive system that does a good job at maximising strength and simplicity, and that is not clearly inferior to an alternative system. Further, (ii) our generalization concerning flash and thunder can be derived from the laws described in

[3]For an accessible summary, sympathetic reconsideration, and refinement of the best system account, the reader is referred to Cohen and Callender [2]. Notably, their account is prepared to face a variety of best systems, as Explanation 1 is. It is worth noting, moreover, that Cohen and Callender [2] are concerned with deductive systematizations of our *knowledge* of whatever domains.

(i). More precisely, it can be derived that the generalization holds universally for a system that contains clouds above the ground and an atmosphere of a specific composition. Otherwise, there would not be an inferential path from the flash to the thunder via the electrical discharge. (i) and (ii) imply that the critical generalization concerning flash and thunder is redundant in the sense of Explanation 1.

In a similar vein, it can be shown that the assertion of a regular connection between the firing of B and E (in the above neuron diagram) is redundant in the just explained sense. In light of these observations, an inferential characterization of non-spurious causes almost falls into place: for any non-spurious cause, there must be a forward-directed inferential path to the effect that all generalizations of this path are non-redundant (in the sense of Explanation 1). Let us express this condition in terms of Ramsey Test conditionals:

Definition 4. $H \vdash_{F-N} C$
$H \vdash_{F-N} C$ iff there is a forward-directed natural deduction proof of C from H such that all generalizations of this proof are non-redundant.

This gives rise to another conditional:

$$A \gg_{F-N} C \in K(S) \text{ iff } (K(S) \div B(A) \vee B(C)), A \vdash_{F-N} C. \qquad (SRT_{F-N})$$

Thus, we obtain:

$$C \gg_{F-N} E \qquad (C1^{**})$$

This condition says that there must be a forward-directed inferential path between cause and effect such that all generalizations of this path are non-redundant.[4]

4 The Final Account

We can now merge $(C1^{*})$ and $(C2^{**})$ into our preliminary account of causal explanation from Section 2.4:

Definition 5. Causal Explanation
Let S be an epistemic state that is represented by a belief base. The set A of antecedent conditions and the set G of generalizations *causally explain* the fact F - relative to S - iff

[4]There is work underway by the authors, showing that $(C1^{*})$ and $(C2^{**})$ are surprisingly powerful in dealing with problems of overdetermination and preemption.

(E1) For all $\alpha \in A$, all $\gamma \in G$, and all $\beta \in F$: $\alpha, \gamma, \beta \in K(S)$.

(E2) For all non-empty $A' \subseteq A$, $\bigwedge A' \gg_{F-N} \bigwedge F \in K(S)$.

(E3) For all $\alpha \in A$ and all $\beta \in F$, $t(\alpha) < t(\beta)$ or $c(\alpha) <_t c(\beta)$.

(E4) For any $\gamma \in G$, (E2) fails to hold for $K(S) \div \gamma$.

If the sets A, G, and F satisfy these conditions, we say that $A \cup G$ stands to F in the relation of *logical entailment with a causal meaning*.

References

[1] Andreas, H. and Günther, M. (2018). On the Ramsey Test Analysis of 'Because'. *Erkenntnis* doi:onlinefirst. URL https://doi.org/10.1007/s10670-018-0006-8.

[2] Cohen, J. and Callender, C. (2009). A better best system account of lawhood. *Philosophical Studies* **145**(1): 1–34.

[3] Gärdenfors, P. (1988). *Knowledge in Flux*. Cambridge, MA: MIT Press.

[4] Halpern, J. Y. and Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science* **56**(4): 843–887.

[5] Halpern, J. Y. and Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *British Journal for the Philosophy of Science* **56**(4): 889–911.

[6] Hansson, S. O. (1999). *A Textbook of Belief Dynamics. Theory Change and Database Updating*. Dordrecht: Kluwer.

[7] Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In *Scientific Explanation*, edited by P. Kitcher and W. Salmon, Minneapolis: University of Minnesota Press. 410–505.

[8] Paul, L. A. and Hall, N. (2013). *Causation: A User's Guide*. Oxford.

[9] Rott, H. (1986). Ifs, Though, and Because. *Erkenntnis* **25**(3): 345–370.

[10] Schurz, G. (2014). Criteria of Theoreticity: Bridging Statement and Non-Statement View. *Erkenntnis* **79**(S8): 1–25.

[11] Spohn, W. (2006). Causation: An Alternative. *British Journal for the Philosophy of Science* **57**(1): 93–119.

[12] Stalnaker, R. (1968). A Theory of Conditionals. In *Studies in Logical Theory (American Philosophical Quarterly Monograph Series)*, edited by N. Rescher, no. 2, Oxford: Blackwell. 98–112.

[13] Strevens, M. (2004). The Causal and Unification Approaches to Explanation Unified—Causally. *Noûs* **38**(1): 154–176.

[14] Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press.

[15] Woodward, J. (2003). *Making Things Happen : A Theory of Causal Explanation*. Oxford: Oxford University Press.

Anti-anti-vaxx: the fairness-based obligation to defer to the expert consensus

Stephen John, sdj22@cam.ac.uk, DRAFT

The aim of this paper is to outline an account of the proper relationship between scientific consensus, on the one hand, and non-experts' beliefs on the other. Before addressing that issue, however, it is useful to start with a refresher course in ethical theory. We often think that there are things which it would be ethically preferable for people to do: for example, to smile at their neighbours, to cut down their carbon emissions, or to not spread deadly diseases. Many theorists argue that there is an important distinction within this set of ethically preferable actions: some actions are decent, but not obligatory (for example, smiling at your neighbours), whereas others are not merely decent, but obligatory (for example, not spreading deadly diseases); and some cases are contestable (for example, cutting down on carbon emissions).[1] Non-performance of the latter kind of action seems to involve violating others' rights or failing in our basic moral duties in a way in which non-performance of the former does not. In turn, it seems that, all else being equal, we can be compelled to perform the second kind of action – or punished for non-performance – but not the first. In principle at least, we might permissibly quarantine people with deadly diseases whereas even the biggest fans of politeness don't think that we can imprison grumpy neighbours.

With this backdrop, let me now turn to the main topic of this paper: the relationship between scientific expert consensus and non-expert belief. As we are all aware, there are various claims which are the subject of a strong scientific consensus, but which are not believed by a significant number of non-scientists; for example, that anthropogenic climate change is occurring, that certain vaccines are safe, or that life has developed via a process of evolution

[1] For a classic discussion of these issues, see Judith Jarvis Thomson, 'A Defence of Abortion' in Kuhse, Schüklenk and Singer (eds.) *Bioethics: An Anthology*. Blackwell, 3rd Edition 2015

by natural selection.[2] I will assume that claims which are subject to a strong scientific consensus are more likely to be true than claims which are not (I return to some complications below). As such, this divergence suggests that some in our community are routinely failing to believe what they ought – in an epistemic sense – believe. Nonetheless, there are often very good, liberal reasons to tolerate such divergence; if someone believes, for example, that humans were created by God, she does not thereby necessarily pose a threat of harm to others. However, sometimes divergences can be "socially relevant", in the sense that failure to believe the relevant claims may have important consequences for others; my belief that some vaccine is unsafe may lead me not to vaccinate my children, posing a threat to your children. If so, we have some legitimate interest in wanting non-experts' beliefs about socially relevant claims to match the scientific consensus: deference would be "ethically preferable".

Does the claim that it would be ethically preferable for non-experts to defer to experts on issues of social relevance express a demand of decency or an obligation? In this paper, I sketch a (very tentative) argument that, at least as long as certain conditions hold, it states an obligation, related to our political obligations of fairness. More snappily, I suggest we each have a political obligation to defer to expert testimony. Before going on, it must be stressed that settling this question does not immediately lead to any normative recommendations. We might think that even if you are obliged not to spread deadly disease, it would be all-things-considered wrong (or maybe just inefficient) to quarantine you. So, too, we might think that it would be wrong (or counter-productive) to force you to defer to experts. Furthermore, my claim is not that people who don't defer are ethically blameworthy. We may fail to meet our obligations for reasons which are outside our control; just as badly designed tax laws may make it impossible for us to discharge our obligation to contribute our fair share to the social

[2] For a useful and up-to-date summary of many such cases see Intemann and de Melo-Martin, *The fight against doubt* (Oxford University Press)

good, so, too, badly designed social-epistemic institutions may impair our ability to meet our obligations of epistemic deference. Still, my question is interesting, because it may change how we think about our social-epistemic environment. I return to this issue in the conclusion

Preliminaries over, I will approach my question in a roundabout way. First, I set out some reasons why individuals might not defer to experts. Second, I show how familiar frameworks in moral and political philosophy for thinking about some of these reasons, those which involve "free-riding", relate to the epistemic case. In the third section, I clarify my arguments against possible objections. Finally, I consider some possible implications of my arguments.

a. Why not defer

There is widespread consensus among the scientific community that the "triple-vaccine" for measles, mumps and rubella (MMR) is both highly effective - at least, as long as a sufficiently high proportion of the population is vaccinated - and safe. As such, all parents are encouraged to vaccinate their children. However, many parents do not vaccinate their children. One key refrain in analysis of this phenomenon of vaccine-refusal is that parents are concerned that the vaccine might cause autism.[3] This concern stems from the work of Andrew Wakefield, a British doctor who published a paper in the *Lancet* suggesting that there *might* be a causal link between the vaccine and autism.[4] The ins-and-outs of the scientific issues here are complex (and further complicated by their relationship to various problems in publication ethics). However, it is safe to say that the vast majority of the scientific community think that Wakefield is wrong; there is a strong scientific consensus that the triple vaccine does not cause autism. Many analyses of the MMR controversy view it,

[3] For a (partial, but useful) overview of these issues,, see See T. Boyce, Health, Risk and News: the MMR Vaccine and the Media (New York: Peter Lang); R. Horton, MMR: Science and Fiction (London: Granta); M. Fitzpatrick, MMR and Autism (London: Routledge).
[4] A.J. Wakefield et al., 'Ileal-lymphoid-nodular Hyperplasia, Non-specific Colitis, and Pervasive Developmental Disorder in Children', The Lancet,

then, as a kind of epistemic failure. Some analyses treat vaccine-refusal as a social-epistemic failure; for example, a failure by scientists to communicate properly or a failure by the media to report properly.[5] Other analyses treat the case as a mass individual-level epistemic failure; for example, as the result of a wrong-headed over-estimation by parents of their ability to grasp complex scientific issues.[6]

Of course, it is entirely possible that these explanations are correct. (Indeed, it's also possible that other issues, such as parents' difficulty in taking time off work may influence vaccine-uptake rates.) Still, I want to suggest that the background assumption that vaccine refusal must involve some kind of epistemic failing is problematic. Doing so overlooks (at least) six reasons for which a parent who is adequately informed of the scientific consensus might refuse to vaccinate a child, none of which involve clear epistemic (or practical) irrationality. At the end of this section, I turn to the implications of this fact for the broader normative question of this paper.[7]

The first two reasons are grounded in the familiar distinction between "is" and "ought". First, one might reason that if others vaccinate their children, then there is no reason to vaccinate one's own children: regardless of whether you vaccinate your child, herd immunity will be maintained, so there is no point in taking a day off work to look after a snotty, crying baby. Clearly someone who reasons this way may be wrong in her assumption that others will vaccinate even if she does not; furthermore, her actions may be unethical (more on this later). Nonetheless, such "free-riding" is not necessarily epistemically irrational nor practically

[5] Horton *op cit*
[6] T. Sorell, "Parental Choice and Expert Knowledge in the Debate about MMR and Autism" (in A. Dawson and M. Verweij eds *Ethics, prevention and public health* (Oxford University Press, 2007))
[7] Note that what follows leans heavily on my own work, and far richer and more recent discussion of vaccine hesitancy in the work of Maya Goldenberg. See S John "Expert testimony and epistemological free-riding" *Philosophical Quarterly* 61(244), July 2011, 495-517, and Goldenberg, Maya J. "Public misunderstanding of science? Reframing the problem of vaccine hesitancy." *Perspectives on Science* 24, no. 5 (2016): 552-581.

irrational.[8] Second, more generally, one may object to the vaccination on other grounds; for example, one might have religious objections to the use of animal products in the vaccine. Again, such reasons not to vaccinate raise difficult ethical issues, but they are clearly consistent with believing, with the consensus, that the vaccine is safe.

The other reasons for vaccine-refusal, by contrast, do involve some kind of refusal to accept the claim that the vaccine is safe as a premise in further reasoning. The third reason rests on the distinction between "population-level" and "individual-level" knowledge. Imagine a parent who reasons as follows: “clearly, there is still some possibility that the vaccine will have some side-effects. Therefore, when the scientists claim that the vaccine is safe, strictly they are making a claim about the overall likely balance of benefit and harm associated with taking the vaccine. That is to say, their claim is that the *average* individual is better-off taking the vaccine than not. I concur with that claim. However, I have some reason to believe that my son is particularly susceptible to suffering some side effect. As such, I suspect that *he* stands to lose more than he gains from the vaccine”. Of course, this parent could be wrong in his belief that his son is “special”. Still, the general pattern of reasoning is not obviously epistemically irrational: it is entirely possible that some medical intervention might be in the interests of the “average” individual in the population, but not in the interests of a specific individual. As such, it might be epistemically rational to refuse to treat a claim about the average case as applying to a specific case.[9]

The fourth and fifth reasons for vaccine hesitancy relate to the epistemology of trust.

[8] For in-depth discussion of “free-riding” in vaccination cases, see A Dawson (2007) ‘Herd protection as a public good: vaccination and our obligations to others’ in Dawson, A. & Verweij, M. (eds.) *Ethics, Prevention and Public Health* (Oxford University Press, 2007).

[9] Goldenberg, 564-566 discusses these issues in great detail. For some of the more general issues here see Fuller, Jonathan, and Luis J. Flores. "The Risk GP Model: The standard model of prediction in medicine." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 54 (2015): 49-61.

Reason four concerns what I call "folk philosophy of science mistrust". I assume that deference to the claims of some expert community depends on two assumptions: first, a sociological assumption, that the community is organised in such a way that members of that community are likely to make (or agree on) claims only when those claims meet certain epistemic standards; second, an epistemological assumption, that these epistemic standards are such that the non-expert should accept the relevant claims.[10] (To see why these two can come apart, note that I might happily admit that the community of astrologers is organised such that they only assert claims when they meet the "epistemic standards" of astrology, but think those standards should not govern my own beliefs).

Let us assume, then, that some parents hold the second assumption about the epistemic standards of biomedical science. They don't, for example, share the worry about population-to-individual inferences, but think that, in general, they should defer to claims which are established relative to the norms of biomedical science. Still, it is entirely possible that they might not trust the *actual* community of biomedical scientists, because they believe that the *actual* community is not set-up such that its members make claims only when they meet the relevant epistemic standards. Indeed, they might have this concern *even if* the community is in-fact well-ordered. This can occur when the non-experts hold what I call a "folk philosophy of science" – i.e. an account of how scientific communities should be ordered – which differs from the correct account of how scientific communities should be ordered. In our case, for example, imagine that a parent believes that a well-ordered scientific community is driven solely by a concern for the truth, regardless of consequences, and is characterised by a high degree of receptivity to the views of mavericks. She then reads of the treatment of Wakefield, apparently drummed out of the community for his heterodox views, and notices that a key

[10] For more on this model of trust, see John, S. (2018). Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty. *Social Epistemology*, *32*(2), 75-87.

concern of many leading epidemiologists seemed to be that Wakefield's work threatened continued control of measles (a concern which is entirely independent of the truth of Wakefield's claims). Faced with this evidence, the parent might decide that the actual community differs systematically from the ideal community, and, as such, refuses to defer to the consensus. Of course, we might think that her "folk philosophy of science" is false.[11] Nonetheless, given the idealised and misleading views of science to which we are all exposed, it is easy enough to see how she might have formed that view. I suggest, then, that it is (at the very least) unclear that her failure to defer to the experts is straightforwardly epistemically irrational; rather, it seems a good example of careful epistemic practice.

One obvious response to this "folk philosophy of science mistrust" is that there can be reasons to trust which do not rest on a model of how communities *ought* to operate, but, instead, on the track-record: I may have no clue how a bicycle workshop "should" operate, but the workshop's past record of successfully fixing my bike may give me good reasons to trust them this time around.[12] Similarly, then, we might argue that, even if non-experts have a false "folk philosophy of science", they also have strong inductive evidence that the biomedical research community should be trusted. Several authors have recently pointed out, we may have reasons to think that research communities are untrustworthy in the sense that the research agenda is skewed in various ways; in our case, for example, towards identifying biochemical treatments for disease, rather than the underlying social and political determinants of health. These arguments are interesting and important, but a wish that the experts had studied a different topic doesn't obviously justify a refusal to defer to what the experts say about the topics they have studied.[13]

[11] For reasons to think that there is no general account of the proper place of dissent in science see Inteman and de Melo-Martin *op cit*

[12] See A.Goldman, 'Experts: Which ones should you trust?" in his *Pathways to knowledge* (OUP, 2002) for an in-depth account of the possibilities of using track-record to assess expertise

[13] (Note that this is one point where I depart from Goldenberg: as I read her analysis of the MMR case, one key problem was a lack of responsiveness from scientists and policy-makers to public concerns. I agree that such a

Even with this caveat in place, note that the appeal to “track record” reasons for trust clearly relies on an assumption: that the experts do, in fact, have a good track record. It is, unfortunately, unclear how good the track record is in our case. There are, after all, some cases – routinely referenced by anti-vaxx communities – where the biomedical research community was wrong. In the UK, for example, vaccine scepticism is often motivated by memories of the thalidomide scandal.[14] It is not, then, contrary to the evidence to think that there is a chance that the community is wrong again in this case, and, as such, to withhold trust.

One might worry about the epistemic status of such meta-inductions: the fact that the medical establishment has been wrong in the past is consistent with it being very unlikely that it is wrong in this case. This leads to the trickiest set of problems of all. We are, by now, all familiar with debates over "inductive risk", and, in particular, how the fact that scientists (apparently) take “inductive risks” may leave a role for non-epistemic values in science.[15] Let me zoom out from the details of these debates to note a far broader point they illuminate: that some claim is "well-established" in some circumstances does not imply that it is "well-established" in all circumstances. Rather, our willingness to act on claims which might be wrong does (and should) vary with the non-epistemic costs of different sorts of error. In general, the greater the costs of acting on a false positive, the more evidence we should demand before accepting some claim.

lack of responsiveness is problematic, and may create circumstances of mistrust. However, I am less clear why such concerns would justify a failure to accept the scientists’ claims).

[14] See Boyce *op cit*

[15] See, *inter alia*, Rudner, R. (1953) “The scientist qua scientist makes value judgements” *Philosophy of Science*, 20(1), Douglas, H. (2000). Inductive risk and values in science. *Philosophy of science*, *67*(4), 559-579, Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science Part A*, *40*(1), 92-101.

Consider, then, the vaccine case.[16] Imagine a parent who accepts that claims about vaccine safety and efficacy are well-established, relative to the epistemic risks which scientists do (or should) be willing to tolerate. However, she also reasons that the stakes in her situation – where her own child's health is at risk – are such that she should demand *higher* standards than the scientists; as such, she refuses to accept the scientifically well-established claim. I suggest that, on the face of it, such refusal to defer to the experts is not necessarily epistemically irrational, insofar as the parent does not deny either the evidence or the consensus. (You might think it is practically irrational in the sense that not accepting the claim the vaccine is safe seems tantamount to accepting that it is not safe, and acting on this claim also carries risks. I don't disagree. Still, the core issue here is that we cannot say that the parent has engaged in epistemically irresponsible reasoning).

Although I think that all six of these (philosophically interesting) concerns can be found in actual debates over vaccine safety, I don't know is how important each is. My analysis is, then, entirely compatible with thinking that the actual phenomenon of non-vaccination has much more to do with, say, the media not reporting claims than with parents understanding but not responding to claims.[17] Why, then, are these possible explanations interesting? At the beginning of this paper, I distinguished two ways to think about non-experts' deference to the scientific consensus: as something nice, but optional, or as something obligatory. In this section, I have distinguished between two broad classes of vaccine-denialism: one class involves a willingness to accept what experts say as correct, but a refusal to act on those claims; the second class involves a failure to accept what the experts say in the first place. My first two possible reasons for denialism speak to the first kind of refusal; the last four concern

[16] For a more detailed account of this argument, see John, S (2012) *ibid*
[17] See Goldenberg *op cit* for an attempt to argue in some more detail that actual hesitancy does involve a complex form of reasoning. Note that, for my purposes, the explanation of *actual* hesitancy is less interesting than the *possible* explanation.

the second. These last four considerations, in turn, seem to imply that, even when the expert community is well-functioning, we may fail to defer to experts without violating basic norms of epistemic rationality. As such, they seem to undercut the claim that we are obliged to defer to the experts; refusing to defer is, it seems, consistent with our epistemic obligations. We seem, then, left with the weaker claim that it would be a nice thing for non-experts to defer to experts - like smiling at your neighbours - rather than something obligatory.

b. Obligations to defer

However, I will now argue that even if there is not a straightforward epistemic obligation to defer to the experts, there may still be another sort of obligation, which I will call a "political" obligation, to do so. To explore this option, consider, again, the first possible grounds for vaccine refusal set out above: the hyper-rational parent who reasons that, as long as others do vaccinate their children there is no point in him vaccinating his own children. Of course, such a chain of reasoning might be challenged on epistemic grounds – how can you be certain others will vaccinate their children? – but it is not inherently epistemically problematic. Furthermore, it is not prudentially irrational; indeed, in some sense, such reasoning is a paradigm form of prudential rationality. Nonetheless, there is a familiar argument that refusing to vaccinate your children on these grounds is ethically problematic. In not vaccinating one is enjoying a public good - herd immunity - without paying one's fair share of the costs necessary for maintenance of that good - vaccinating one's children. The prudential anti-vaxxer is "free-riding", and, hence, violating norms of fairness.[18] Assuming that we have a basic obligation to treat others fairly, we have an obligation to vaccinate our children. Of course, this argument has to be handled carefully: for example, there is

[18] For the details of such arguments, see Dawson *op cit*

disagreement over how to analyse obligations of fairness where enjoyment of the public good is non-voluntary, as in the case of herd immunity. Furthermore, any obligation to vaccinate may have to be balanced against other ethical considerations, such as religious freedom, or more general concerns about the proper role of the state in family life. Still, very many political philosophers and bio-ethicists agree that we can (and often do) have fairness-based obligations to contribute to public goods, even when doing so is not in our prudential interests (narrowly construed).

The fairness argument generates an obligation to vaccinate for those parents who *do* defer to the consensus. What, though, can it tell us about whether parents *ought* to defer to the consensus? Consider the following argument:

1. As a matter of fact, there is widespread deference to scientific experts in our community
2. Such widespread deference generates various, non-rival and non-excludable benefits for all members of society: i.e. they generate a public good
3. Each who benefits from a public good has a *prima facie* political obligation of fairness to sustain that good

Therefore, we each have a *prima facie* political obligation to defer to experts

I have just discussed premise 3 in the context of vaccinating one's children. Of course, that premise could be attacked. However, I will simply assume that the consensus view in political philosophy is (broadly) correct. Note that if we cannot assume that premise, we

might have even more serious problems – lack of a decent account of the legitimacy of many State actions – than non-experts' failure to defer to expert testimony.

Premise 1, by contrast, might seem very shaky: after all, the entire concern driving this paper is, precisely, that in some cases non-experts do not defer to experts. However, high-profile cases of non-deference are consistent with a general tendency to defer to expert opinion (particularly, consensus among experts). I claim that there is such a tendency. Consider some examples: we believe the experts who tell us that smoking cigarettes causes lung cancer, that the aeroplanes won't fall out of the sky, and that the weather tomorrow will be sunny. Indeed, it is notable that even those who seek to disrupt the processes by which non-experts learn from experts – so-called, "agnotologists" – often adopt strategies which assume that non-experts do defer to expert consensus: for example, rather than deny that we should defer to experts, they set up their own "experts" and they argue that the apparent consensus within the scientific community is real. This is not a flippant point. Agnotologists are, I suggest, the real experts in social epistemology; unlike philosophers, their livelihood depends on understanding how knowledge actually moves through societies.[19] At a more abstract level, one might argue that such deference is "practically necessary" in modern societies, because such societies require us to use complex forms of knowledge, and are characterised by a division of epistemic labour. Taken together, then, these considerations suggest that Premise 1 may be less controversial than familiar claims about a "crisis of trust" suggest.

To ground a fairness-based obligation for deference, we need to do more than show that each does often defer to experts. Rather, we need to establish the existence of a "public good": that patterns of deference generate some further good we all enjoy, where that enjoyment is both "non-rival" (i.e. me getting more doesn't mean you getting less) and "non-excludable" (i.e. the

[19] For a longer version of this argument, see John 2018, op cit

benefit cannot be "gated off" for a select few). This is not trivial: it is entirely possible that patterns of deference benefit each one of us individually, but not generate a "public good", akin to "herd immunity". In that case, we might each have a good reason to defer to the experts, but nothing like a fairness-based obligation to do so. I suggest, however, that the patterns of deference do generate such a shared good: as a result of such patterns, our lives are both predictable and stable, because we can better predict how others will respond to aspects of our shared social-epistemic environment. A shopkeeper who knows that others will, in general, defer to the expert opinion of weather forecasters can predict she needs to order more umbrellas; a citizen who knows that epidemiologists' advice on the dangers of smoking will be widely believed can assume that others will listen to her arguments for banning smoking in public; we can all make some well-grounded predictions about how others will respond to claims made in public debate. In turn, these goods of predictability and stability are both non-rival (me “taking advantage” of Amy’s predictability doesn’t leave less predictability for you) and non-excludable (I cannot easily prevent you making use of the fact that Amy is predictable).[20]

I do not mean to downplay the range and depth of disagreement in our society. We cannot blithely assume that others will believe what the scientists say and plan our lives accordingly. However, I do suggest that premise 1 and 2 taken together do capture one important aspect of how we live together, and our attitudes towards experts. Therefore, there is an important parallel to be drawn between two arguments: one, familiar in political philosophy and bio-ethics, shows that people who believe that a vaccine is safe are obliged to vaccinate their children; the second, sketched above, shows that people who benefit from widespread

[20] Note an interesting twist here: the purely prudential free-rider case, discussed at the start of this section, clearly involves free-riding on this pattern of epistemic deference. It’s only because the free-rider can assume that others will defer to the experts that she is able both to defer to the experts and to know that there is no point in vaccinating her own child.)

deference to expertise should believe that vaccines are safe. Both arguments appeal to fairness considerations: in both cases, the argument generates a "political" (rather than prudential or epistemic) obligation. Specifically, in the latter case, I have suggested that we have a “fairness-based” obligation to defer to expert testimony; in refusing to defer we are making use of a public good – others’ willingness to defer to the experts – while refusing to “do our share” in maintaining that good. Note that, like the fairness-based obligation to vaccinate one’s children (if one believes that the vaccine is safe), this obligation may need to be balanced against other ethical concerns; it does not imply, for example, that we can simply lock up anti-vaxxers. Still, it seems that we can say that deference would not merely be ethically preferable, but, at least *prima facie* obligatory.

c. The epistemic and the political

Imagine that we live in some perfectly fair and just society. However, this social harmony has arisen only out of a set of social practices centred around some patently false claim, such as that the Universe was created by little demons in 1888. I have strong epistemic reasons to doubt such a claim. Nonetheless, a friend argues with me as follows: “you have an obligation to ensure that our community continues to function harmoniously; as such you have a strong political obligation to continue to believe that the Universe was created by little demons”. Many of us would, I think, find this suggestion patently absurd. I may have some obligation to keep quiet about my beliefs (although even this is arguable), or to state them in a respectful, careful manner. However, it seems odd to say that my political obligations provide me with a reason to believe some epistemically unjustified claim. It might seem in turn that my argument above is prone to the same sort of concern: we cannot have a *political* obligation to *believe* claims. This section explores and addresses this concern.

Broadly, I suggest we can distinguish two sources of objections to my conclusions. First, one might hold that epistemic voluntarism – i.e. the claim that beliefs are subject to our will – is false. Given the "ought implies can" principle, then, it does not make sense to say that we ought to believe a certain class of claim.[21] However, this objection is unconvincing. Even if it is true that we cannot simply "choose" what to believe, it is still possible for us to engineer social situations and to cultivate frames of mind which increase our chances of obtaining certain sorts of beliefs. Of course, such interventions are not guaranteed to succeed, but there is no reason to think that they are bound to fail. Even if epistemic voluntarism is false, we can restate my conclusion as the claim that we have a political obligation to engineer a situation where we acquire certain beliefs or habits of mind. (That is to say, I see no argument against "indirect" voluntarism, even if there is a problem with "direct" voluntarism).

The second objection is more important: one might hold that we can, reasonably, talk of what we "ought" to believe, but hold that such claims depend solely on epistemic reasons, rather than ethical or political reasons. In response to this concern, note that my arguments do not deny that our beliefs should, in large part, be guided by a straightforwardly epistemic reason: that there is consensus within a scientific community. This marks an important difference between my arguments and the "little demons" case: the obligation to defer is not based *solely* on political considerations, because, I assume, the relevant obligation is only operative when the expert community is, in fact, trustworthy. Therefore, my argument is not intended to suggest that what we should belief should be untethered from our evidence. Rather, my claim is that in cases where there are both epistemic reasons in favour of believing some claim and considerations against believing that claim (including other epistemic

[21] For the classic statement of such concerns, see Williams, Bernard. "Deciding to Believe." In *Language, Belief, and Metaphysics*, ed. Howard E. Kiefer and Milton K. Munitz, 95-111. Albany: SUNY Press, 1970

considerations, such as the track record of the relevant epistemic community), we should disregard the second set of reasons, on broadly "political" grounds.[22]

I am not, then, saying that we should believe whatever would be "politically best" for us to believe, regardless of the evidence. What I am claiming, however, is that we should sometimes believe claims, even when we have epistemic reasons not to believe them. That is to say, my conclusions deny that we should enjoy epistemic autonomy. How should we assess this result? Note that any claim about our obligations is a claim that certain forms of autonomy should be limited. For example, an obligation to vaccinate our children limits our parental autonomy. This is simply a re-description of our obligation. It is only a knockdown counter-argument against such an obligation if we assume that parental autonomy should never be curtailed. However, no-one believes that parental autonomy is entirely sacrosanct. Similarly, to object to my argument on the grounds that it limits epistemic autonomy simply raises the question of how much we should value epistemic autonomy. Valuing epistemic autonomy on purely epistemic grounds seems odd; after all, we would often be far better-off epistemically speaking simply deferring to others. It seems, then, that the value of epistemic autonomy must itself be understood in ethical or political terms; say, as necessary for self-development or societal advancement. No doubt those are worthy goals, but as ethical and political goals, they can be balanced against other ethical and political concerns, such as, I suggest, demands of fairness.

Conclusion

[22] Note (to be completed properly later): I can't quite grasp how these comments relate to the huge on-going debates over the proper place of non-epistemic values in science. I am a bit lost here, insofar as the claim is stronger than that we can use "non-epistemic values" to respond in cases of underdetermination, etc, and, yet, doesn't look like standard denials of the "epistemic priority thesis" (insofar as the epistemic retains a kind of priority!)

A recent spate of books and papers has argued that many apparent failures to defer to experts – often framed by the media and commentators in simplistic terms of "scepticism about science" – are, in fact, more complex than they first appear. When we consider the sheer complexity of the ways in which we must judge expertise, and the important – and proper – roles for non-epistemic values in shaping the production, dissemination and reception of scientific knowledge, then we can better understand how non-experts may rationally fail to defer to experts. As Section 1 of this paper showed, I think that such arguments are correct, in the sense that non-experts' refusal to abide by the consensus may be grounded in complex epistemic and ethical norms. However, I want to add a coda: it does not therefore follow that all disagreement is, thereby, to be tolerated or accepted. Disagreement, dissidence and non-compliance from social practices can be, in some sense, reasonable – it really is a bit silly to vaccinate your child if herd immunity obtains – but politically unacceptable. I have argued that certain failures to defer to experts may be like that.

What follows? I don't know. It certainly does not follow, for example, that it would be ethically permissible to lock-up vaccine denialists in Maoist re-education camps. Maybe, once we take the all epistemic, ethical and political considerations into account, all that follows is a reaffirmation of familiar claims about the importance of public education and public debate. However, I do suggest that thinking in terms of obligations may help us reshape how we frame certain debates. Individuals' ability to meet their political obligations is always structured by broader institutions and practices. Sometimes, we cannot help but free-ride, if, for example, we cannot get the day off work to vaccinate our children. One key theme in some recent writing has been whether we can characterise "normatively inappropriate dissent", and, if so, how we should respond. My model proposes a way of thinking about some of these issues: certain forms of dissent are problematic not only because

they have deleterious consequences (both epistemic and non-epistemic), but because they involve engineering social-epistemic situations where citizens cannot meet our political obligations. Thinking in terms of obligations does not solve our problems, but enriches our normative vocabulary for thinking about them.

THE META-INDUCTIVE JUSTIFICATION OF INDUCTION: THE POOL OF STRATEGIES

TOM F. STERKENBURG

Abstract. This paper poses a challenge to Schurz's proposed meta-inductive justification of induction. It is argued that Schurz's argument requires a notion of optimality that can deal with an expanding pool of prediction strategies.

1. Introduction

Schurz (2008; 2009; most recently, 2018; 20xx) proposes a justification of induction based on *meta-induction*, induction at the level of competing methods of inference. The argument proceeds in two steps. First, there is the *analytical* justification of meta-inductive strategies in the setting of sequential prediction. This consists in mathematical results on these strategies' optimality, as established in the machine learning branch of *prediction with expert advice* (see Cesa-Bianchi and Lugosi, 2006; Vovk, 2001). Second, there is the *empirical* observation that *object-induction*, induction at the level of events, has been most successful so far. Hence, the argument goes, the optimal meta-inductive strategy favors the object-inductive strategy, thus justifying it.

Schurz's proposal is a refinement of Reichenbach's attempted *pragmatic justification* or *vindication* of induction (see Salmon, 1967, 52ff, 85ff). The fundamental idea underlying both is that the aim for *reliability*, guaranteed success, may be replaced for *optimality*, guaranteed success *whenever some method would be successful*. This weaker aim is still reasonable, because the cases in which *no* method can be successful are in an obvious sense not so interesting—in those cases there is simply nothing we could do. And, importantly, this weaker aim looks more feasible: while it appears impossible to design a single inductive method that can take into account everything nature could possibly do (this is in a sense the original problem of induction, see Howson, 2000), it looks more feasible to design a single method that tracks *what we could possibly do*. Thus Schurz (2018, 3895) proclaims that "optimality justifications constitute new foundations for foundation-oriented epistemology."

The obvious qualm is whether the aim of a truly general optimality is really more feasible. This qualm finds a sharp expression in the question *what class of methods* we should actually require optimality *for*. In this paper, I investigate this question within the context of Schurz's argument. My conclusion will be that the argument needs an optimality that covers *expanding* pools of strategies, which suggests that things may not be easier, after all.

The plan of the paper is as follows. First, I will briefly describe the presupposed framework of sequential prediction (sect. 2) and the structure of Schurz's argument

Date: October 30, 2018.

(sect. 3). (This is based on a much more detailed reconstruction of the argument elsewhere, Sterkenburg, 20xx.) In order to constitute an actual justification for object-induction, the conclusion of the argument also needs us to accept that the *optimality* of the meta-inductive strategy amounts to a *justification* for it. The question whether this is really so then prompts us to have a closer look at the pool of prediction strategies assumed.

I start with the objection due to Arnold (2010) that the optimality results that Schurz relies on are restricted to finite pools of strategies (sect. 4). I point out why Schurz's argument, to go through at all, *must* presuppose a finite pool; but I argue that it does not *need* an infinite pool to yield the desired justification: it only needs optimality relative to the (necessarily finitely many) actually proposed alternatives. However, I then argue (sect. 5) that this does involve something more: it needs a notion of optimality that is robust against *new* strategies being proposed over time.

2. The framework of prediction

2.1. **The framework of sequential prediction.** We define a prediction game as a triple $(\boldsymbol{y}^\omega, \Pi, \ell)$ of a *history* $\boldsymbol{y}^\omega$, a pool Π of *prediction strategies*, and a *loss function* ℓ.

A *history* $\boldsymbol{y}^\omega$ is an infinite sequence of *events*. Events are identified with values in some set Val of possible values. Write y_n for the n-th element of $\boldsymbol{y}^\omega$, or the event in *round* n of the game ($n \in \mathbb{N}^{>0}$).

Predictions are elements in some set $\mathrm{Val}_{\mathrm{pred}}$. A *prediction strategy* P, an element of the pool Π, specifies in each round n a prediction $\mathrm{pred}_n(P)$ about the next event.

In this paper, I will restrict attention to the central class of *probabilistic prediction games*. In these games we assume binary events, $\mathrm{Val} = \{0, 1\}$, and predictions that are probabilities (for the next event being 1, say), $\mathrm{Val}_{\mathrm{pred}} = [0, 1]$.

Strategy P, when making prediction $\mathrm{pred}_n(P) = \mathrm{pred} \in \mathrm{Val}_{\mathrm{pred}}$ for round n, suffers, when the outcome is revealed to be $y_n \in \mathrm{Val}$, a certain *loss* $\ell(\mathrm{pred}, y_n)$. That is, a loss function $\ell : \mathrm{Val}_{\mathrm{pred}} \times \mathrm{Val} \to [0, \infty)$ quantifies how much a prediction was off in light of the actual outcome.

A basic example is the *absolute* loss function, defined by $\ell_{\mathrm{abs}}(\mathrm{pred}, y) = |\mathrm{pred} - y|$. Another loss function, prominent, among other things, for its strong connection to *Bayesian* prediction (sect. 3.1 below), is the *logarithmic* or *log*-loss function defined by

$$\ell_{\log}(\mathrm{pred}, y) = \begin{cases} -\ln(1 - \mathrm{pred}) & \text{if } y = 0 \\ -\ln \mathrm{pred} & \text{if } y = 1 \end{cases}.$$

For given loss function, the *cumulative* loss of P by the conclusion of round n is the sum $\mathrm{Loss}_n(P) := \sum_{i=1}^{n} \ell(\mathrm{pred}_n(P), y_n)$. The *loss rate* $\mathrm{loss}_n(P)$ of P by n is the average $\mathrm{Loss}_n(P)/n$ of its losses up to n.

2.2. **The goal: an optimal strategy.** Given a pool Π of prediction strategies, we aim to design a *meta-inductive* strategy MI that, having access to the predictions of all the other strategies, predicts in such a way that it is *optimal* with respect to Π. That is, by following MI we will *always* do about as good as, in hindsight, we possibly could have done—given that the strategies in Π represent what we could have done. Here 'always' means: on *every* single history of events.

What it means for a meta-inductive strategy to be 'about as successful' as any other strategy we make precise in terms of the divergence between MI's loss rate

and the quantity $\text{minloss}_n := \min_{P\in\Pi\cup\{\text{MI}\}} \text{loss}_n(P)$, the minimum loss rate among all the strategies (including MI itself) by round n. Specifically, we seek a function f, that depends on n and inevitably also on the size $K := |\Pi|$ of the pool of strategies, such that for all rounds n,

$$\text{loss}_n(\text{wMI}) \leq \text{minloss}_n + f(n, K). \tag{1}$$

A minimal requirement is that f is such that it entails *long-run convergence*,

$$\lim_{n\to\infty} (\text{loss}_n(\text{MI}) - \text{minloss}_n) = 0, \tag{2}$$

for which it at least needs to be decreasing in n. But as we will see below, there actually exist prediction algorithms that achieve bounds (1) for f that decrease in n at a very fast rate, giving strong *short-run* guarantees.

3. The argument

3.1. **Step one: the analytical optimality of meta-induction.** A general type of meta-inductive strategy is the *weighted-average* strategy waMI, specified by

$$\text{pred}_{n+1}(\text{waMI}) := \sum_{P\in\Pi} w_n(P) \cdot \text{pred}_{n+1}(P). \tag{3}$$

Here the *weight function* w_n assigns a weight to each strategy P based on its past success.

An important example of a weighted-average strategy in the probabilistic binary prediction game, for the particular choice of the log-loss function, is the *Bayesian strategy* BayMI, that updates its weights via Bayes's rule. It is given by

$$w_n(P) = \frac{w_0(P) \cdot \exp(-\text{Loss}_n(P))}{Z}, \tag{4}$$

with normalization term $Z = \sum_{P\in\Pi} w_0(P)\cdot\exp(-\text{Loss}_n(P))$. Here w_0 is some prior probability assignment or *initial weight function* over Π. With a *uniform* initial weight assignment, where $w_0(P) = 1/K$ for each $P \in \Pi$, assignment (5) simplifies to

$$w_n(P) = \frac{\exp(-\text{Loss}(P))}{Z}, \tag{5}$$

so that the weights depend on the strategies' performance only.

Now one can derive that BayMI, for the log-loss function, satisfies, for each $P \in \Pi$,

$$\text{Loss}_n(\text{BayMI}) \leq -\ln w_0(P) + \text{Loss}_n(P). \tag{6}$$

Choosing again a uniform w_0, this translates in the short-run optimality bound

$$\text{loss}_n(\text{BayMI}) \leq \text{minloss}_n + \frac{\ln K}{n}. \tag{7}$$

That is, for this game we can achieve bound (1) with f of order $1/n$. What is more, it turns out to be possible, for a wider class of loss functions, to design strategies that explicitly mimic the Bayesian strategy for the log-loss function, for *these* loss functions, in order to achieve a similar bound. Thus for these so-called *mixable* loss functions, which include the quadratic loss function, there also exist meta-inductive strategies with bounds of order $1/n$. These are the strongest possible bounds for any game; but for an even wider class of loss functions, that also includes the absolute loss function, it is still possible to define meta-inductive strategies—specifically,

exponentially-weighted strategies that can also be seen as generalizations of the Bayesian strategy—with bounds of order $1/\sqrt{n}$.

Taking stock, we have that for a wide class of games there exist meta-inductive strategies that are optimal in a very strong sense. Moreover, these optimal strategies predict by combining weighted predictions of all the other strategies in the pool, where the weights depend on these strategies' attractiveness or past performance—and in the case of uniform weights, on their past performance *only*. In particular, the strategy in the pool that so far has been performing best receives the largest weight: it is in that sense that we say that the meta-inductive strategy *favors* the most successful strategies so far. Thus the first step of Schurz's argument is that

(A) The meta-inductive strategy MI, that at each point in time favors strategies to the extent of their relative success so far, is an optimal method.

3.2. **Step two: the empirical success of object-induction.** The second step is the empirical observation that "so far object-induction has turned out to be the most successful prediction strategy" (Schurz, 2008, 304).

In (20xx), I argued that the relevant perspective here is to view the object-inductive or *scientific* method as competing with a number of alternative *non*scientific methods. Importantly, for Schurz's argument it is not necessary to further specify what this scientific method actually consists in, the notorious problem of description (see, e.g., Lipton, 2004). It is enough to recognize that there is something like the scientific procedure, that we wish to find justification for; and, plausibly, that its predictions have been highly successful so far, at least more successful than those of nonscientific alternatives. Thus the second step of Schurz's argument is that

(E) As a matter of empirical fact, the object-inductive strategy OI, that we identify with the scientific method (and that we imagine to be in competition with various proposed nonscientific methods), has been, at this point in time, the most successful prediction strategy (among the pool Π of all of these competing strategies).

3.3. **Conclusion: meta-induction favors object-induction.** From (A) and (E) it follows that

(C) The meta-inductive strategy MI for the pool Π of OI and its nonscientific competing strategies, an optimal strategy for Π, favors most, at this point in time, the object-inductive strategy OI.

In (20xx), I noted that, for (C) to yield the desired justification of OI, we also need to say that an optimal strategy favoring OI actually amounts to a justification for it. The discussion of this step brings out an important limitation of the argument: it cannot provide a justification for the object-inductive *strategy* (for *always* sticking with object-induction), but at best—though this would still be an important result—a justification for sticking with the object-inductive prediction *for now* (thus allowing for the possibility that in the more distant future it will no longer be a good strategy to follow).

Furthermore (ibid.), I noted that we would also still need to argue that the optimality of the meta-inductive strategy actually amounts to a justification for following it. I allowed that the notion of optimality is sufficiently strong that it does—*given* that the pool of strategies is appropriate, a proper rendition of all we could possibly do. We will now investigate whether this is truly so.

4. The restriction to a finite pool

Arnold (2010) points out that the analytical justification of meta-induction does not extend to pools of *infinitely* many strategies, and suggests that this is a problem for Schurz's proposal.

4.1. **The impossibility result.** Arnold's observation, his impossibility theorem 3 (ibid., 589), comes down to the following. For every strategy MI we might propose, nature can construct an adversarial history that makes it fail maximally: in each round, it can choose $y = 0$ precisely if our strategy's pred > 0.5. Then our strategy's total loss grows linearly, and its loss rate $\text{loss}_n(\text{MI})$ never goes to 0. However, for a rich enough infinite pool of strategies, say a pool that includes all *computable* strategies, there exists for *every* finite history (including every finite initial segment of the adversarial history we are constructing), some strategy that has managed to predict this history *perfectly*. That is, $\text{minloss}_n = 0$ for every round n. Hence our strategy's loss rate does not converge to the best strategy's.

This shows that optimality is impossible to achieve in the general case of infinite pools of strategies: at least for sufficiently rich such pools, we can for any given strategy construct a history that refutes its optimality.

4.2. **Universal but non-uniform optimality.** Arnold writes (ibid., 592, emphasis mine), "If only a finite number of prediction strategies are taken into account, then we exclude the overwhelming majority of *possible* prediction strategies from the game right from the beginning." Arnold's suggestion that Schurz's argument needs a notion of optimality relative to infinite pools of strategies thus appears to be motivated by a more definite demand: the argument would need a notion of optimality relative to the infinite pool of *all possible strategies*. The argument would need an optimality that is no longer relative but truly *universal*.

The general move from reliability to optimality actually makes universality look genuinely more feasible. Namely, it seems reasonable to "take into consideration only those prediction strategies that can be described by an algorithm" (ibid.; see Sterkenburg, 2018 for more details). While there seems little justification for limiting possible histories to computable sequences of events, it does seem reasonable to limit the methods of prediction we could possibly devise to the computable ones. Rather than the continuum of all possible histories, we then only need to consider the vastly more restricted pool of computable prediction strategies.

Of course, this is also still a countably infinite number, and so optimality in the original sense is ruled out by the impossiblity result above. However, we can still attain a weaker, *non-uniform* optimality for countably infinite pools.

Consider again the Bayesian strategy in the log-loss game, and the bound (6) on its cumulative loss: notably, this bound holds just as well for an initial weight assignment over a countably *infinite* pool of strategies. Thus even in case of an infinite pool Π, we can still derive, parallel to (7), that for every strategy $P \in \Pi$, for all n,

$$\text{loss}_n(\text{BayMI}) - \text{loss}_n(P) \leq \frac{-\ln w_0(P)}{n}, \tag{8}$$

so that in particular, for all $P \in \Pi$,

$$\lim_{n\to\infty} (\text{loss}_n(\text{BayMI}) - \text{loss}_n(P)) \leq 0. \tag{9}$$

The crux here is that (8) depends on the given predictor, specifically, on the initial weight $w_0(P)$ assigned to it. Now in case of only a *finite* number of strategies, it is possible to *uniformly* assign each the *same* initial weight, and we can derive a uniform bound (1), and consequently uniform convergence (2). But in case of an infinite number of strategies, this is obviously impossible: we are forced to give some non-uniform initial weight assignment. Consequently, the convergence (9) is non-uniform: we are *not* guaranteed to eventually match the success of all strategies in the pool, *at the same time*. We are guaranteed, for any given strategy in Π, to eventually match the success of this strategy, but by the time we do other strategies might always still be way ahead of us; this is the reason why this bound is consistent with the impossibility result of sect. 4.1 above.

But it is still something—given a pool Π, and in the absence of stronger guarantees, one can argue there is some justification for sticking to a non-uniformly optimal strategy, for sticking to a strategy that is guaranteed for any selected strategy from Π to eventually match this strategy's success. Granted this, there is certainly some justification for sticking to a strategy that is guaranteed for *any possible strategy* to eventually match this strategy's success—for a *universally* non-uniformly optimal strategy.

Now if we identify all possible strategies with the computable ones, then the Bayesian strategy over all computable strategies, that is non-uniformly optimal relative to all computable strategies, would be universally non-uniformly optimal. Could this then be a strategy that meets Arnold's demand?

Unfortunately, it cannot, and the reason is that this Bayesian meta-inductive strategy is no longer computable itself. This follows from a diagonal argument that goes back to Putnam (1963), an impossibility argument that is actually very similar to that of sect. 4.1 above. What it means is that, on our earlier restriction of the possible strategies to the computable ones, the candidate optimal strategy is actually no longer a proper strategy; nor is any optimal strategy for the pool of computable strategies. This quandary holds with great generality: it is not restricted to the log-loss function, and we cannot escape it by looking for weaker computability constraints (Sterkenburg, 2018). Thus Arnold's demand is, indeed, unrealizable, even on a weaker notion of non-uniform optimality: there cannot be a universally optimal prediction strategy.

4.3. **Infinite pools in Schurz's argument.** On reading Arnold's presentation, one gets the impression that Schurz's proposed justification of induction boils down to the description of an optimal strategy. In contrast to "[m]ost of the proposed solutions to the problem of induction [that] tried to prove the reliability of the inductive procedure," he writes, "Schurz, following Reichenbach, merely tries to show the optimality of a specific inductive strategy" (2010, 585). With the understanding that this must be *universal* optimality, such a project, we just discussed, is indeed doomed to fail.

But Schurz's actual argument is more subtle than that. As explained in sect. 3 above, the argument seeks to justify object-induction, a strategy that is presumably not optimal itself. The meta-inductive strategy, optimal relative to all of OI's competitors, only comes in to confer justification to OI. Now even if one insists that OI's competitors are *all possible* strategies, things might still look better for Schurz's actual argument—perhaps, for instance, it is not so important here that a universally optimal strategy cannot actually be a proper (i.e., computable) strategy

itself? But we can save ourselves the trouble of going into this: unfortunately, there is a more direct reason why the pool of competitors *must* be finite, for Schurz's actual argument to work.

To see this, we return to the observation in sect. 4.2 above that the optimality bound (7), and in general a bound (3) for any meta-inductive strategy, must involve an initial weight assignment w_0. It is only with a uniform prior, which is only possible for a finite pool, that the initial weights all cancel out and P's weights in later rounds depend on its success *only*. Thus for a countably infinite pool of strategies, a meta-inductive strategy *must* express some prior preference for some strategies above others, that works through in the posterior weights.

But this is devastating to Schurz's argument. Conclusion (C) follows from step (E) if the optimal strategy at this point in time favors the most successful strategy, OI. In case of a finite pool of strategies, where the weights are only determined by the success, it does. But in case of an infinite pool, the meta-inductive strategy only favors OI at this point of time *if it assigned strategy* OI *a sufficiently high initial weight*. The meta-inductivist will *not* favor OI, even if OI has been the most successful strategy, if it assigned OI too low an initial weight (and conversely, it *would* favor OI, even if OI had *not* been successful at all, if this were compensated by a high enough initial weight). In short, the meta-inductive justification of object-induction would have to presuposse a sufficiently strong prior preference for object-induction, and this would render it an obviously circular argument.

4.4. **The finite pool of actually proposed alternatives.** Thus in the end Arnold is right to worry that Schurz's argument is not compatible with an infinite pool of competing strategies: indeed it is not. This leads us to "the philosophical question whether an optimality result demonstrated for a finite number of prediction strategies might suffice to answer the problem of induction" (ibid., 585).

Schurz writes, "I make the realistic assumption that [the meta-inductive strategy] has finite computational means, whence I restrict my investigation to prediction games with finitely many strategies" (2009, 206; also see 2008, 284). More precisely, Schurz (2018, 3891) offers in defense of the limitation to finite pools an

> *Argument from cognitive finiteness*: Epistemic subjects are assumed to be finite beings. Finite beings can simultaneously access (and compare) only finitely many methods of finite complexity. Therefore the optimality justification of meta-induction is not affected by the finiteness restriction.

The second statement is not strictly true, though, and anyway does not entail the desired conclusion. It is not strictly true, because finite computational means are consistent with weighing over an infinite enumeration of strategies. (We can plausibly only give probabilistic—real-valued—predictions up to some finite accuracy, and since we also have to give decreasing weights to the strategies in the enumeration, there are in each round only finitely many strategies that can have an impact; yet this is different from stipulating a finite pool from the start.) But more importantly, as Arnold noted already, the observation that a meta-inductive strategy can only deal with finitely many strategies falls short of a justification for this restriction: in itself, this "merely amounts to admitting that under this 'realistic assumption' [the meta-inductive strategy] simply cannot always perform optimally" (2010, 592).

Arnold continues, "[a]s there is no logical contradiction involved in the assumption of an infinite number of alternative strategies, the only grounds on which it could be defended are empirical" (ibid.). These are exactly the grounds, I will now argue, on which it *can* be defended, in the context of Schurz's actual argument.

Again, Schurz's argument is not to identify a universally optimal strategy, optimal among the infinity of all possible strategies; it is to justify object-induction, from the empirical observation (E) that object-induction has been most successful so far. Most successful among what? Certainly *not* among all possible strategies—we can probably conceive, in hindsight, of strategies that would have been more successful still. No: object-induction has been most successful, so far, among *all actually proposed alternative strategies*. The relevant empirical observation (E) is that object-induction has been most successful among the various actually proposed nonscientific strategies—of necessity a *finite* number of strategies.

It is in this sense that Schurz (2018, 3891) is surely right when he, after his initial and unconvincing defense, adds that "[i]n any case, the problem of choosing among finitely many competing methods captures the most important part of the induction problem." Now the problem is to give a good reason for sticking to OI, rather than turning to one of its contestant strategies; and the hope, again, is to derive such a noncircular reason with the help of the optimality of a meta-inductive strategy, that by (E) favors it. But then it seems enough to have an optimality relative to *this same pool* of all actually proposed strategies. The pool of all actually proposed strategies seems to properly represent all we could have done, and so an optimal strategy for this pool would be justified.

Unfortunately, there is still a crucial sense in which this optimality falls short of including all we could have done.

5. The restriction to a fixed pool

The basic intuition, again, behind the optimality of the meta-inductivist over the pool of all proposed strategies, is the one going back to Reichenbach: for every possible history, and for every alternative strategy proposed, if this strategy is successful, the meta-inductivist will mimic it and be successful, too. Thus Schurz (2008, 304) concludes by once more evoking this intuition to answer the obvious skeptical reservation: "how can it *ever* be possible to prove that a strategy is optimal with respect to *every* other accessible strategy in *every* possible world—without assuming anything about the nature of alternative strategies and possible worlds?" To understand how this is possible, Schurz answers, one should note that "meta-induction has an unlimited learning ability: whenever this strategy is confronted with a so far better method, it will learn from it and reproduce its success" (2018, 3892).

There is, however, a clear sense in which *this is not true*: namely, when the meta-inductivist is confronted with a *new* strategy.

5.1. **The expanding pool of actually proposed alternatives.**

A meta-inductivist can be optimal for a finite pool of strategies, like the finite pool of actually proposed strategies, but, crucially, we need to assume that this pool is *fixed*. Yet it is only plausible that the pool of actually proposed strategies will *expand* in the course of time: informed by the actual history of events, brand *new* strategies may be proposed.

The meta-inductive strategies we have been considering cannot guarantee optimality with respect to new strategies—simply because they do not allow for dynamically incorporating new strategies in their pool. Imagine that we fix the pool of strategies that have been proposed by this time in history, and design and follow a meta-inductive strategy that is optimal relative to this pool. But in the future a new strategy might be proposed, and this strategy might continue to be forever much more successful than all the original strategies—and hence than our meta-inductive strategy. This means that our meta-inductive strategy is no longer optimal in the sense of being as good as we can possibly be: surely we could have followed the new and much more successful strategy instead.

5.2. **Truly analytical optimality.** Is this really a problem for Schurz's argument, though? Was the goal, specified above, not to justify following object-induction among the alternative strategies we have *now*?

Yes, this is still the goal, and the relevant empirical fact (E) is still that object-induction has been most successful among the alternatives we have *now*. However, I now claim, the *analytical* step (A), to be truly analytical, must involve a notion of optimality that is robust against all possible empirical circumstances: against all possible histories of events, but also against *all possible evolutions of the pool of strategies*.

Again, the crucial component of analytical optimality is that it covers every possible history: it should not and does not depend on the contingent fact of the actual history of events we have seen occur. But likewise, it should not depend on the contingent fact of the actual alternative strategies that we have seen proposed. This does *not* mean that we must demand optimality relative to all possible finite pools of alternative strategies *at the same time* (this would be Arnold's infeasible demand of optimality relative to all possible strategies); but it does mean that we must demand optimality relative to all possible *expanding pools*, or histories of finite pools.

Otherwise, the meta-inductive method is simply not optimal in the sense of analytically the best we could do. The meta-inductive method that we fix at this point of time, relative to the current pool of alternative strategies, was not guaranteed to be optimal: it might not have been if other, better, strategies had been proposed in the past. And it might still fail to be, if other, better, strategies are proposed in the future. As such, it is not a strategy one is justified to follow without any empirical assumptions, and it cannot fulfill the analytical role required for Schurz's argument.

5.3. **Dynamic optimality.** What are the prospects for the design of a 'dynamic' meta-inductive strategy that *is* optimal in the above sense?

Such a strategy must allow for dynamically adding new strategies to its pool as they appear, while somehow preserving optimality guarantees with respect to all the available strategies in each round. There are some choices to be made here, starting with a suitable standard of optimality.

It appears too strict, for instance, to demand that the meta-inductivist keep its loss rate low with respect to new strategies on *past data*: it cannot, of course, guard itself against new strategies that simply fit their past predictions to the past data and thereby can claim to have a perfect score. This demand indeed goes beyond a notion of optimality as the best we could do, since we could only have followed

a strategy from the moment it is actually available. On the other hand, it also appears infeasible to first measure the success of a new strategy from the moment it comes in. As an extreme scenario: in each round a new strategy appears that makes an initial perfect prediction and then stops predicting well; now after each round the best strategy again has a perfect score while the meta-inductivist has not necessarily been doing very well.

So perhaps we need to argue for a middle way, where new strategies are assigned some 'virtual' loss for the rounds where they were not yet participating. This is indeed an approach taken in the literature that comes closest to our problem, the framework of 'specialists,' experts that are in each round allowed to 'sleep' and refrain from making predictions (Freund et al., 1997). The 'abstention trick' due to Chernov and Vovk (2009) advocates the assignment to asleep strategies of the same predictions as the meta-inductivist; using this trick Mourtada and Maillard (2017) derive bounds for the specific case of growing expert pools. However, it remains to be argued that these results are truly applicable to the current context: that they can still support both the analytical and the empirical step of the argument. I will leave this here at this briefest of sketches, and suggest further investigation as a challenge for Schurz's research programme.

6. Conclusion

I identified as a challenge for Schurz's proposed meta-inductive justification of induction the need for a notion of optimality that is robust against newly proposed prediction strategies. Notably, this challenge finds a parallel in the problem of new theory in the traditional Bayesian framework (Earman, 1992, 195ff; Gillies, 2001). This suggests that the aims of reliability and of optimality are confronted with much the same structural difficulties, and that, unless this challenge can indeed be met, a shift of focus to optimality might not be such an effective means of avoiding foundational problems as Schurz advocates.

References

E. Arnold. Can the best-alternative justification solve Hume's problem? On the limits of a promising approach. *Philosophy of Science*, 77(4):584–593, 2010.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, 2006.

A. Chernov and V. G. Vovk. Prediction with expert evaluators' advice. In R. Gavaldà, G. Lugosi, T. Zeugmann, and S. Zilles, editors, *Proceedings of the 20th International Conference on Algorithmic Learning Theory*, volume 5809 of *Lecture Notes in Computer Science*, pages 8–22. Springer, 2009.

J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. A Bradford Book. MIT Press, Cambridge, MA, 1992.

Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th Annual ACM Symposion on the Theory of Computing*, pages 334–343, New York, 1997. ACM Press.

D. A. Gillies. Bayesianism and the fixity of the theoretical framework. In D. Corfield and J. Williamson, editors, *Foundations of Bayesianism*, volume 24 of *Applied Logic Series*, pages 363–379. Springer, 2001.

C. Howson. *Hume's Problem: Induction and the Justification of Belief*. Oxford University Press, New York, 2000.

P. Lipton. *Inference to the Best Explanation*. Routledge, London, second edition, 2004.

J. Mourtada and O.-A. Maillard. Efficient tracking of a growing number of experts. In S. Hanneke and L. Reyzin, editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 517–539. PMLR, 2017.

H. Putnam. 'Degree of confirmation' and inductive logic. In P. A. Schilpp, editor, *The Philosophy of Rudolf Carnap*, volume XI of *The Library of Living Philosophers*, pages 761–783. Open Court, LaSalle, IL, 1963.

W. C. Salmon. *The Foundations of Scientific Inference*. University of Pittsburgh Press, Pittsburgh, PA, 1967.

G. Schurz. The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science*, 75(3):278–305, 2008.

G. Schurz. Meta-induction and social epistemology: computer simulations of prediction games. *Episteme*, 6(2):200–220, 2009.

G. Schurz. Optimality justifications: new foundations for foundation-oriented epistemology. *Synthese*, 195(9):3877–3897, 2018.

G. Schurz. *Hume's Problem Solved: The Optimality of Meta-Induction*. 20xx. Manuscript in preparation.

T. F. Sterkenburg. *Universal Prediction: A Philosophical Investigation*. PhD Dissertation, University of Groningen, 2018.

T. F. Sterkenburg. The meta-inductive justification of induction. To appear in *Episteme*, 20xx.

V. G. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Munich Center for Mathematical Philosohy, LMU Munich
E-mail address: tom.sterkenburg@lmu.de

Intervention as both Test and Exploration: Reexamining the PaJaMo Experiment based on Aims and Modes of Interventions

Hsiao-Fan Yeh, Adjunct Assistant Professor, Department of Philosophy, National Chung Cheng University, Taiwan.
Ruey-Lin Chen, Professor, Department of Philosophy, National Chung Cheng University, Taiwan.

Abstract

This paper explores multiple experimental interventions in molecular biology. By "multiple," we mean that molecular biologists often use different modes of experimental interventions in a series of experiments for one and the same subject. In performing such a series of experiment, scientists may use different modes of interventions to realize plural goals such as testing given hypotheses and exploring novel phenomena. In order to illustrate this claim, we develop a framework of multiple modes of experimental interventions to analyze a series of experiments for a single subject. Our argument begins with a brief characterization of Craver and Darden's taxonomy of experiments, because the taxonomy they have made implies various modes of interventions (Carver and Darden 2013). We propose to extract two interventional directions and two interventional effects from their taxonomy as the basis of classification. The vertical or inter-level direction means that an intervention is performed between different levels of organization and the horizontal or inter-stage direction means that an intervention is performed between different stages of a mechanism. Interventions may produce an excitatory or an inhibitory effect. As a consequence, we can classify modes of interventions according to different directions and effects. We illustrate our claims by doing a case study of the PaJaMo experiment, which is a series of experiments for a single subject. The final goal in this paper is to provide a taxonomy of characteristics of experimentation in which the PaJaMo experiment is adequately located.

Keywords: Intervention, Exploration, Experimentation, the PaJaMo experiment, Scientific Practice

1. Introduction

The molecularization of biology has raised a range of new questions and issues for the philosophy of biology, especially questions about the uses of experiments and interventions. For example, how interventional experiments contribute to discovering new phenomena and adding new knowledge in molecular biology, what experimentally causal reasoning is used in identifying mechanistic components that are responsible for some phenomenon, whether or not there are different kinds of interventions can be discerned, and how many kinds of interventions utilized by molecular biologists. These questions are important, because molecular biology is basically an experimental science and uses interventional method in almost all of experiments. More particularly, molecular biologists often use a series of organized experiments for exploring a single subject. By performing that series of organized experiments, they realize plural aims and attain plural outcomes. How do molecular biologists design and organize a series of experiments for a single subject? In order to answer these questions, the best way is to examine an actual complicated experiment in the history of molecular biology.

The PaJaMo experiment performed by Arthur Pardee, Francois Jacob, and Jacque Monod, is one of the most famous experiments in the history of biology (Craver and Darden 2013:138). Philosophers of biology such as Kenneth Schaffner (1974a, 1974b, 1993) and Marcel Weber (2005) examine it from the perspective of theory generation; and Carl Craver and Lindley Darden (2013) also analyze the experiment from the perspective of testing and discovering schemas of mechanisms.[1] Their contributions offer much help in our understanding of this famous experiment. In this paper, we would like to add a new understanding from the perspective of experimental interventions for realizing both testing and exploration aims.

Testing is a recognized aim of making experiments. In the end of the 20th century, exploratory uses of experimentation have been explored (Burian 1997; Steinle 1997). Philosophers tends to make a dichotomous distinction between testing (or theory-driven in Steinle's term) experimentation and exploratory experimentation. However, we wonder whether or not there is a kind of experiments that are both testative and exploratory. Waters (2004, 2008) provides a non-theory-centric methodology which combining explanatory reasoning and investigative strategies for classical and molecular genetics. Although Waters did not analyze experimental interventions in details, his work can serve well as an exemplar for us to develop a similar analysis on the interventions in the PajaMo experiment.

[1] Darden and Craver have drawn attention to the discovery of a new and key component in the mechanism of protein synthesis in the PaJaMo experiment (Darden and Craver 2002). On their account, the discovery of messenger RNA (presumed to be ribosomal templates) required the integration of aspects of the mechanism by interfiled. Molecular biologists studied forward from the DNA to the next stage in protein synthesis while biochemists worked backward from peptide bonds to activated amino acids (Darden and Craver 2002:80-84). However, they don't analyze the PaJaMo experiment in details nor focus on the notion of experimental interventions.

Mechanistic philosophers such as James Woodward (2002)[2], Craver (2007), and Craver and Darden (2013) have widely analyze experimental interventions, connecting them with the analyses of mechanisms. From the view of new mechanical philosophy, knowledge of mechanisms is necessary for understanding, predicting, and controlling biological phenomenon (Machamer, Darden, and Craver 2000; Darden and Craver 2002; Glennan 2002; Bechtel and Abrahamsen 2005; Darden 2006). Experimental interventions are used by molecular biologist as powerful instruments to help produce knowledge of biological mechanisms (Machamer, Darden, and Craver 2000:17). Carver and Darden (2013) further distinguish various types of experiments in which a taxonomy of interventional modes is implied. Inspired by the pioneering philosophers' work, we intend to make the implicit taxonomy of interventional modes explicit and to inquire whether or not different modes might be used together in a series of organized experiments for both test and exploration.

A taxonomic framework of modes of experimental interventions is characterized by the following three points: (i) We can distinguish different modes of experimental interventions according to two standards: the interventional *direction* and the interventional *effect*. (ii) Two interventional directions (vertical/inter-level and horizontal/inter-stage) and two interventional effects (excitatory/positive and inhibitory/negative) can be identified. The vertical or inter-level direction means that an intervention is performed between different levels of organization and the horizontal or inter-stage direction means that an intervention is performed between different stages of a process. (iii) In a series of related experiments, scientists can use multiple interventional modes to *test* given hypotheses and to *explore* novel objects.

Since our study is inspired by the classification of experiments in Craver and Darden (2013), we will begin by summarizing their new mechanistic philosophy and offer a brief characterization of their classification of types of experiments in section 2. Section 3 reinterprets the classification as a framework of interventional modes. Then, we argue that the reinterpretation provides a new framework of the modes of experimental interventions. In section 4, we introduce a new kind of experimental interventions, the inter-stage kind, which Craver and Darden do not mention. Section 3 and 4 jointly argue for the points (i) and (ii). Section 5 takes the PaJaMo experiment on the synthesis of β-galactosidase in *E. coli* to illustrate all the points (i)-(iii). The final section provides a taxonomy of characteristics of experimentation in which the PaJaMo experiment is adequately located.

2. Craver and Darden's classification of types of experiments

Ever since 2000, Craver and Darden jointly develop a mechanism-based and dis-

[2] Woodward has proposed a counterfactual account of the concept of mechanism (Woodward 2002). He argues that components of mechanisms should behave in accord with regularities that are invariant under interventions. On his account, Jacob and Monod's lac operon model and the experimental results will be correctly described and predicted by the notion of modularity of mechanism. Melinda Fagan also has analyzed lac operon model (Fagan 2016). By contrast with Woodward, she argues that the mechanical account with interventions will not always provide an entirely satisfactory account for the case of double preventions or omissions. She proposes a new, complementary mechanistic explanation for that case. However, they focalize on the operon model rather than the PaJaMo experiment in details. The PaJaMo experiment is an important base of construction of the theoretical model.

covery-oriented methodology for biological sciences, especially molecular biology and neuroscience (Machamer, Darden and Craver 2000; Craver and Darden 2001, 2005; Darden and Craver 2002). They majorly analyze reasoning strategies that are used for discovering mechanisms that underling living phenomena. In addition to these analyses, they contribute a long chapter in their 2013 book, *In Search of Mechanisms,* to analyze how experimentation works to help discover mechanisms (Craver and Darden 2013, Ch. 8).

Craver and Darden argue that discoveries of mechanisms are usually made piecemeal via repetitive refinements, which can be guided by interventional experiments. They also analyze processes in which scientists use experimental interventions to test schemas (or models) of mechanisms and then to discover actual mechanisms. In search of a full mechanism, scientists may manipulate some part of a mechanism, intervene in its process, and then observe the changes occur in the termination condition of the mechanism. The observed changes provide a piece of useful evidence or a guide for discerning the entity and activity that are causally relevant to the behavior of the mechanism from that are not. Scientists use the information from manipulations and interventions to infer what could come before or next in the mechanism by "backward chaining" or "forward chaining"[3] (Darden and Craver 2002). A whole picture of the mechanism is thus puzzled out. In their words, scientists use interventional experiments to transform a how-possible constructed schema into a how-actual description of a mechanism (Machamer, Darden and Craver 2000:17; Craver and Darden 2001; 2005:235; Darden and Craver 2002).

In order to provide a full analysis of experimentation, Craver and Darden make a taxonomy of experiments in Chapter 8 of *In Search of Mechanisms*. They distinguish loosely three categories of experiments: those for testing causal relevance, those (interlevel experiments) for testing componential relevance, and those (complex experiments) for asking specific mechanistic questions. The second category is classified into three subkinds: interference experiments that are bottom-up and inhibitory, stimulation experiments that are bottom-up and excitatory, and activation experiments that are top-down and excitatory. The third category is in turn categorized into three subcategories: by-what-activity experiments, by-what-entity experiments, and series of experiment with multiple interventions. They also discuss the famous PaJaMo experiment under the independent title "preparing the experimental system," showing that the experiment also uses multiple interventions.

Craver and Darden emphasize that their goal "is not to offer a systematic taxonomy of experimental types but rather to call attention to the ways that experiments…, to answer specific questions about how a mechanism works." (Craver and Darden 2013:119) However, their work still leaves a strong impression that they are making a taxonomic system of experiments, not only because they use the term "kind" in the context but also because they classify kinds into subkinds. Moreover, they say that interlevel experiments have "the three most common kinds" (p.126) and consider

[3] According to Darden and Craver, backward and forward chaining are reciprocal strategies for discovering mechanisms. When scientists reason about one part of a mechanism on the basis of what is already known in the schematic mechanisms, they can reason from the beginning by forward chaining or from the end by backward chaining. Forward chaining use the experimental results in early stages to reason or conjecture about the information that are likely to be found in later stages; backward chaining is just the reverse (Darden and Craver 2002).

"some alternative kinds of experiments" that fail to fit their intervene-and-detect structure (p.129). All indicates that they are classifying kinds of experiments in the framework of new mechanical philosophy.

Two features in Craver and Darden's taxonomy of experiments are noteworthy. The first feature is that their taxonomy implies a taxonomy of interventional modes, which will be examined in next section. This implication allows us to interpret and treat their taxonomy of experiments as a taxonomy of interventional modes. The second feature is that Carver and Darden pay more attention to experimental tests in the process of discovering a mechanism while say less about experimental investigation, exploration, and discovery. However, experiments in molecular biology often perform many functions other than testing.

Exploratory uses of experimentation have been gotten attention since the end of the 20th century (Burian 1997; Steinle 1997). Exploratory experiments are "driven by the elementary desire to obtain empirical regularities and to find proper concepts" and "typically takes place in those periods of scientific development in which – for what ever reasons – no well-formed theory or even no conceptual framework is available or regarded as reliable." (Steinle 1997: S70) Moreover, a few philosophers of science note that many experiments in classical and molecular biology share the characteristics of exploratory experimentation (Waters 2004; Burian 2007; O'Mallye 2007). Other philosophers such as Kenneth Waters (2008) and Chen (2013) argue that, in classical and molecular genetics, incidental discoveries of novel phenomena in the process of experimenting can be used as investigative tools to discover mechanisms and construct hypotheses or models. For examples, Gregor Mendel's hybridization experiment with peas incidentally discovered the segregation and the independent assortment of hereditary units and led to the discovery of Mendelian mechanism of heredity. Frederick Griffith's experiment with *Pneumococcus* discovered the transformation of bacteria cells and led to a series of discoveries of molecular mechanisms of heredity (Chen 2013). Molecular biologists, M. Hammarlund, E. Jorgensen, and M. Bastianis, learned the crucial guidelines from the unexpected phenomena in their experiment and lead to the discovery of the function of β-spectrin protein in neurons (Waters 2008). In those cases, scientists incidentally discovered the novel phenomena by performing an experiment or a series of organized experiments without the direction of theories, and those discoveries in turn urged them to search for the underlying mechanisms. Experiments that discovered those novel phenomena use interventions and possess an exploratory or investigative characteristic. The previous discussion indicates that reconsidering the exploratory or investigative function of experimental interventions will shed light on the analysis of the PaJaMo experiment and other similar ones in molecular biology.

3. Modes of Experimental Interventions

Craver and Darden (2013) classify the second category of interlevel experiments into three subkinds based on their so-called the invervene-and-detect structure. Interference experiments are bottom-up and inhibitory. Stimulation experiments are bottom-up and excitatory. Activation experiments are top-down and excitatory. In those bottom-up experiments, "one intervenes into a component in a mechanism and detects

changes in the behavior of the mechanism as a whole." (Craver and Darden 2013:125-126) In those top-down experiments, one intervenes on the start conditions to manipulate the phenomenon and detects the behavior of the components in the mechanism. However, we wonder whether or not there are top-down and inhibitory experiments, one intervenes on the start conditions to inactivate or inhibit the phenomenon and detects the behavior of putative components in the mechanism. We think that vaccination experiments are the very kind. In a vaccination experiment, scientists inactivate or reduce the pathogenicity of some kind of pathogenic bacteria by physical and chemical methods or kill them, and then inject the attenuated or killed bacteria vaccine into subjects, and see if target organs of subjects no longer manifest relevant symptoms.

Since all "subkinds" of interlevel experiments share the intervene-and-detect structure, we should interpret the four experimental types as four interventional modes. We view bottom-up and top-down as two different *interventional directions* that are not mutual exclusive, because one may exert different interventional directions into the same mechanism. Similarly, we view excitation and inhibition as two *interventional effects* that are neither mutual exclusive, because an intervention may produce both excitatory effect on one component and inhibitory effects on another component in the same mechanism. Thus, we build up a temporary framework of experimental interventions that has four modes based on the two directions and the two effects: a top-down excitatory, a top-down inhibitory, a bottom-up excitatory, and a bottom-up inhibitory intervention. This is the distinction of interventional modes rather than the taxonomy of experimental types, because different modes may be applied in one and the same experiment.

Consider Julius Axelord's series of experiments that Craver and Darden use to exemplify their third category of series of experiments with multiple interventions. In the series of experiments, Axelrod and his colleagues seek to discover the mechanism for regulating neurotransmitters by conducting multiple interventions at different stage. Craver and Darden reports:

> First, he injected rats with norepinephrine to increase their blood pressure….killing the nerves innervating the eyes and the salivary glands…In a second intervention they then injected the cats with labeled norepinephrine….In a third intervention, they then stimulated the live sympathetic nerve and showed that…transmitter is in fact released from the neurons. In a fourth intervention…they showed that they could prevent the labeled transmitter from being re-sequestered by treating the nerves with cocaine. (Craver and Darden 2013:134-137).

Two features are worth to be pointed out. First, these interventions are operated in both the *top-down* direction on killing the nerves and the *bottom-up* direction on injecting norepinephrine and stimulating the live nerves. They are not mutual exclusive. Second, these interventions produce both an *inhibitory* effect and an *excitatory* effect that are neither mutually exclusive. For example, injecting the neurons with cocaine brings about an inhibitory effect for the nervous system on blocking the reuptake of neurotransmitters, but also produces an *excitatory* effect for endogenous neurotrans-

mitters between two neuron synapses on remaining the positive effect. Interventional effects might be usually opposite, but not be mutually exclusive necessarily. It depends on how initial state or base line being changed by intervention. An excitatory effect should trigger or excite the behavior of the mechanism while an inhibitory effect should eliminate or shut-down the phenomenon. Axelrod's series of experiment just shows that the two interventional directions may be combined with the two interventional effects to form the common framework for the categories of "experiments for testing causal relevance".

4. "Inter-stage" interventions

We now further argue that the framework can be adequately applied to the category of "experiments for testing componential relevance." If our argument is right, then Craver and Darden's distinction between "experiments for testing causal relevance" and "interval experiments for test componential relevance" is unnecessary. Consider vaccination experiments. If some kind of pathogenic bacteria in a vaccination experiment is confirmed as the *etiological cause* of the relevant symptoms and disease, then the interlevel experiments for testing componential relevance can be also interpreted as the experiments for testing causal relevance if we take an integrated view of causality and mechanisms which is developed in Darden (2013:24-26)and Chen (2017:141-143).[4] According to this view, a mechanism has at least five causal aspects: (1) a mechanism as a complete cause, (2) a mechanism piece as a partial cause, (3) a stage in a mechanistic process as a partial or complete cause in a causal chain, (4) an activity in a mechanism as a cause of micro-change in the mechanism, and (5) a disturbance as a cause of an abnormal output of the mechanism (see Fig. 1). These aspects include part-whole relations, for example, entities and activities are parts of mechanism or lower-level mechanisms are parts of higher-level mechanisms; and cause-effect relations, for example, a previous stage is a cause of later stage in a mechanism or making a difference by intervention is a cause of what happens in the later stage.

[4] Darden have articulated the temporal feature of mechanisms. What she argues is that the interfeild relation between Mendelian and molecular biology best characterized as "investigating different, serially integrated, hereditary mechanisms" (Darden 2005). Mechanisms of the two fields operate at different times and are composed of different working entities of different sizes. Our interest in this paper is exploring another aspect of temporal feature, that how a stage of a given mechanism is identified to an etiological factor, and what mode of this intervention will be.

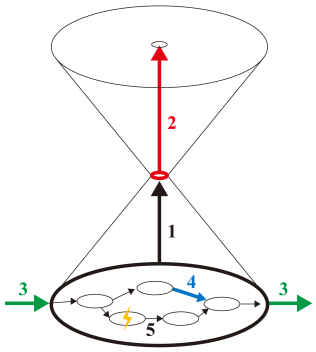


Figure 1. Integrating five causal aspects into a hierarchical structure. Arrow 1, 2, and 4 represent part-whole relations; arrow 3 and lighting bolt with arrow 5 represent cause-effect relations.

This integrated view of mechanism and causality allow us to introduce a kind of "inter-stage" interventions in order to build a more complete framework of experimental interventions. For this purpose, we especially focus on (3), because, since biology is going more and more molecular, even genomic and post-genomic, molecular scientists often need to design and operate interventional experiments to investigate continuous stages of the mechanism involved in embryology, epigenetics and developmental biology. For example, in order to understand how cells having identical genomes develop differentiated and transmit particular characters to the offspring, scientists need to produce various mutations as interventional tools and then to compare of target DNA sequences and the relative gene regulation and expression in a hierarchy of mechanism levels. Their goal is to discover the various stages occurring in a particular order of a mechanism.

That's why we need to take the temporal and etiological factor into our account. In such cases, the causal relevance exists between different stages rather than between different levels. As a consequence, we have "inter-stage" experimental interventions, whose direction is horizontal. By contrast, the direction of the bottom-up and top-down interventions is vertical.

Interventions in the vertical direction occur between two levels, for examples, neuroscientists may put the rat in a maze and record the electrical activity of neurons in the rat hippocampus, and molecular biologists may intervene on one nucleotide sequence that codes some genetic information in organisms and observe the impact in the behavior of a mechanism as a whole. As we have seen in Section 3, vertical or "inter-level" interventions have the "top-down" and the "bottom-up" kinds or modes. Interventions in the horizontal direction occur in different stages in a mechanistic process, for example, molecular biologists may engineer a part of a mechanism at the phase of the initiation of transcription and investigate changes on the later steps of elongation or termination. Particularly the regulatory mechanisms have different working entities serially operating at different times in an extended process. The inter-level kind of interventions can be used to test and investigate any putative part of

some mechanistic model while the inter-stage kind can be used to test and investigate any putative stage of some mechanistic process.

To determine the feature of an entity or activity involved in the mechanism, scientists attempt to *make some difference* at the inputs and see whether such an intervention brings about some corresponding change at the outputs. If the entity or activity plays an excitatory role in the mechanism (such as an excitatory neurotransmitter and promoter), then removing it should induce or in some way prevent the phenomenon. If the entity or activity plays an inhibitory or regulating role in the mechanism (such as an inhibitory neurotransmitter or repressor), then removing it should excite or at any rate change the phenomenon. If, in contrast, the part plays no role in the mechanism, then making a difference of the component should be of no consequence. There are typically two kinds of consequences: excitatory effect and inhibitory effect. We call the consequences that are tending to activate, excite, stimulate the original states are "excitatory effects". Those that are tending to eliminate, shut-down, inhibit the original states are "inhibitory effects". Adding the horizontal modes to the previous four modes, a we have six interventional modes: top-down excitatory, top-down inhibitory, bottom-up excitatory, bottom-up inhibitory, inter-stage excitatory, and inter-stage inhibitory.

In the excitatory kind of inter-level mode, one intervenes on the start level to enhance or activate some part (e.g., injecting norepinephrine or stimulating the live nerves) and observes the reactions on another level. In the inhibitory kind of inter-level mode, one intervenes on the start level to weaken or inhibit some part (e.g., making a mutation on some sequence of genes) and detects the reactions on another level. In the excitatory kind of inter-stage mode, one intervenes to trigger or increase transcription of the regulated gene at the upstream stage (e.g., adding an activator to help polymerase binding the promoter) and see the changes at the downstream stage. In the inhibitory kind of inter-stage mode, one intervenes to inactivate or shut down a part in the upstream stage and assess the changes at the downstream stage of a particular gene (e.g., knocking out a gene in an organism and investigating the effect of gene loss).

Here we extend Craver and Darden's intervene-and-detect structure to a fuller framework of experimental interventions. We distinguish six interventional modes according to three directions and two kinds of effects. Furthermore, we will argue that the six modes can occur in a single experiment or a series of organized experiments for one and the same subject next section. This work not only answers the question about how many modes of experimental interventions used by molecular biologists, but also gives a more comprehensive view to the role of experimentation contributing to acquire new biological knowledge.

Given a new framework of experimental interventions with six modes, we want to analyze how different interventional modes to be used to test hypotheses and to investigate novel phenomena, entities, or objects.

A basic aim of experimentation is to test hypotheses. There are different types of hypotheses to be put into tests, for examples, a causal hypothesis, or a mechanistic model as a whole, or a putative part (an entity or an activity) of a mechanistic model, or a putative stage of a mechanistic model. In order to uncover a complete mechanism underlying a phenomenon, scientists may perform a lot experiments or a series of ex-

periments to test all causal hypotheses related to a mechanistic model and all assumptions of putative parts and stages of the mechanistic model. Thus, we have classification of tested targets (hypotheses) rather than that of experiments. Different targets may be the common goal of one and the same series of experiments.

Consider the second basic aim of experimentation: to investigate or to explore novel objects. There are many different types of objects to be investigated or discovered, for examples, a new significant phenomenon, or a new entity, or a new kind of activity, or a new mechanism. In order to puzzle out and uncover a complete mechanism underlying a discovered novel phenomenon, scientists may need to perform a series of experiments that can investigate all working entities and all relevant kinds of activities. Thus, we have classification of discovered objects rather that of experiments, because a series of experiments may be performed to discover all relevant objects.

Experimental targets and objects are not mutual exclusive, neither are interventional directions and effects. They all may occur in one and the same experiment or a series of experiment for a single subject. A series of experimental interventions may be performed for two experimental aims, use two interventional directions, and acquire two interventional effects. One can exert an intervention into some mechanism and produce excitatory effects on one component and inhibitory effects on another in the same mechanism, depending on whether or not the feature of the intervened entity is essentially excitatory or inhibitory. Interventions may produce novel or unexpected phenomena that are used to test a mechanistic model and discover the mechanism as a whole. All these experimentally interventional modes make important discoveries in biology. The PaJaMo experiment is the best example for illustrating our claim.

5. Experiments with multiple modes of interventions: discovering the synthesis of β-galactosidas

For a long time, historians of science have characterized the contribution of the PaJaMo experiment to the advancement in molecular biology and have argued the question that who should get the credited with the discovery of the repressor model (Morange 1998). Philosophers of biology have been more concern the questions of how the scientists generated and justified the repressor hypotheses and how the related experiments contributed to discovery mechanisms.

Kenneth Schaffner is the first philosopher who analyzed the PaJaMo experiment from the perspective of theory generation. Under the influence of the logical empiricism, he argues that there is a "unitary logic" covering the reasoning of discovery and justification in the generation of the repressor hypotheses (Schaffner 1974a, 1974b, 1993).[5] On his account, one does not need two kinds of generative contexts or reasoning pattern to understand the generation of new theories.

Marcel Weber (2005) criticized Schaffner's claim that the reasoning employed in generation of new theories is the same as that in justification. He argues that the gen-

[5] Logical empiricists distinguish between the "context of discovery" and the "context of justification". According to the traditional view, philosophy of science only concerns the way in which new theories are corroborated or justified while the way in which new theories are constructed or discover is the subject for history, psychology, or sociology of science (Popper 1959).

uine generative reasoning in the PaJaMo experiment is a kind of analogical reasoning. "In my view, the crucial question on which the validity of Schaffner's conclusion turns is the following: Granted that the repressor model follows deductively from the complete results of the PaJaMo experiment plus some background assumptions, does this deductive argument reflect how Jacob and Monod *actually* generated this hypothesis? Monod's own recollections quoted above suggest that this is not the case. Rather, what Monod implies is that the repressor hypothesis was generated by an *argument from analogy*." (Weber 2005:62) Although analogical reasoning cannot be derived from a deductive argument nor be used in logical justification, it still occurs in the generation of new theories. Weber argues that analogical reasoning can be a kind of rational reasoning pattern even though it does not employ any general rules or procedures (Weber 2005: 55-63). In order to support the argument, Weber draws attention to the PaJaMo experiment. However, he analyzes the experiment in the subject of how biologists generate new theories or hypotheses by solving problems. He does not focus on the methodology of the experimentation in details.

Craver and Darden focus on the experimental analyzing. They propose a mechanism-discovering approach to analyze how the PaJaMo experiments contributes to discover new mechanistic schemas (2013). They take the PaJaMo experiment as a searchlight to reveal how series of experiment with multiple interventions contributes to the construction of new mechanistic model.

> This experiment…ushered in an entirely new way of thinking about the mechanisms by which organisms regulate gene expression. Yet it would distort the structure of the experiment to see it an instance of an experiment for testing causal relevance…or as an experiment driven by the goals of identifying components in a mechanism…Rather, the significance of the experiment lies in its ability to test a hypothesis about the active organization of the mechanism, to reveal that mechanism involves the inhibition of an inhibitor. (Craver and Darden 2013:140)

On their account, the PaJaMo experiment is a series of experiments involving multiple interventions. However, their work is providing the taxonomy of experimental types and their methodology is partial to the mechanism-centered approach. Based on their work, we develop a new framework of interventional modes to reexamine the series of PaJaMo experiments. The new framework comprises plural experimental aims and interventional modes. Among them, the inter-stage intervention as one of the interventional modes has not been mentioned by philosophers of biology before. Our goal is to provide a more complete view for generation of new biological theories that goes beyond the older dichotomies: the distinction between justification and discovery and between theory-centered and phenomenon-centered. In what follows, we first introduce the background, unsolved problems, given hypotheses, and how new phenomenon discovered in the PaJaMo experiments in order.

5.1 Background and unsolved problem

Ever since the middle of the twentieth century, molecular biologists studied the

phenomenon of "diauxy" by combining genetic and biochemical approaches. The puzzling phenomenon was that when two food sources (say, glucose and lactose) were given in a microbial culture, the *Escherichia coli* bacteria (hereinafter to be referred as bacteria) digested one type of food sources (glucose) first, after a latency period, it digested the other type of food sources (lactose). By many experiments, biologist observed that the phenomenon showed two successive growth curves and separated by a period of lag (see Fag. 2).

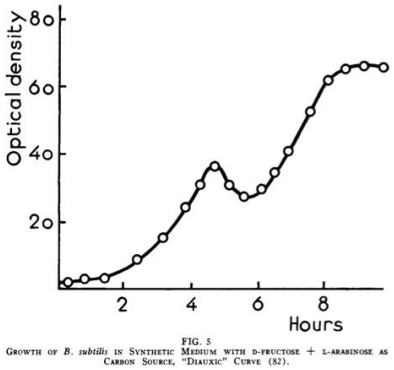


Figure 2. The phenomenon of diauxy was that bacteria would display different digest ability corresponding to specific food resource. It showed two growth curves and separated by a period of lag. Reproduced from Monod, Jacques (1947). "The phenomenon of enzymatic adaptation: And its bearing on problem of genetics and cellular differentiation." p. 251.

The so-called phenomenon of diauxy was first noticed around 1900. In the 1930s and 40s, biologists viewed it a kind of adaptation. They thought that bacteria first produced a kind of enzyme to digest glucose, then, *developed* the ability to produce another kind of enzyme to digest lactose. They called the phenomenon "enzymatic *adaptation*," because bacteria needed some moments to produce different enzyme. However, this term was dropped for its teleological meaning later.

Another conjecture for the phenomenon was that bacteria contained a general enzyme which could take on various properties for different circumstance changes. When glucose was present, the general enzyme displayed one shape to digest it; when lactose was present, the general enzyme switched another appropriate shape to digest it. The phenomenon of diauxy looked like that lactose stimulated or *induced* bacteria to making the corresponding changes. In the 1950s, biologists renamed the phenomenon "enzyme *induction*" Lactose was called "inducer" because biologists believed that lactose induced bacteria to produce enzymes. However, they did not really understand that what the nature of the "inducibility" was and how the mechanism underling the phenomenon operated.

Today we know that the mechanism underlying the phenomenon of induction is gene regulation, which can be simply described as the following: when both glucose and lactose are present in a microbial culture, bacteria will digest glucose first; when glucose is consumed, the genes of bacteria will begin to synthesize a specific enzyme

(say, β-galactosidase) to digest lactose. In other words, only when lactose is present, the genes synthesize β-galactosidase.

Since 1950s, biologists discovered several main genes involved in the mechanism of gene regulation. The genes can be distinguished to two kinds: structural genes and regulatory genes. There are three structural genes: *lacZ*, *lacY*, and *lacA*. The *lacZ* gene is responsible for encoding the β-galactosidase enzyme, which cleaves lactose into glucose and galactose, both of them are utilized as energy source for the cells. The *lacY* gene is responsible for encoding the enzyme permease, which inserts into cell membranes and transports lactose into the cells. The *lacA* gene is responsible for encoding β-galactoside transacetylase. These three structural genes are expressed only when lactose is present and glucose is absent. The regulatory genes are responsible for encoding two kinds of proteins: an activator called catabolite activation protein (CAP) and a repressor. The CAP normally binds to a specific site on DNA at or near the promoter (which is a region of DNA that initiates the transcription of structural genes) and the repressor binds to the operator (which is a region that regulates the activity of genes). The repressor is encoded by the *lacI* gene. When glucose is present, the repressor will repress the transcription. On the contrary, when glucose is absent and only lactose is present (or an external inducer is added), the repressor will be displaced from operator by inducers, namely repression of the repressor, thus, the *lacZ* and *lacY* genes are *de-repressed* (say, expressed). In this way, the combined effect of two regulators, the activation of the CAP and the repression of the repressor, will make the structural genes to be expressed. So what the unsolved problem here is that how did scientists discover that the enzyme induction could be effected by a mechanism of the repression of a repressor.

5.2 Hypothesis to be tested

The enzyme induction (lactose induces bacteria to produce the enzymes) appears to be a kind of the positive activities of some entity. The repression of the repressor appears to be a kind of the negative activities of some entity. How did an ostensibly positive activity be discovered that was caused by the effect of double negative activities? How did the scientists generate a new mechanistic model for explaining the phenomenon?

In the preparatory stage of the PaJaMo experiment, the scientists only knew that the *lacZ, lacY,* and *lacI* genes played important roles in enzyme induction. But they did not know that how these genes interacted with each other and how these genes organized altogether. They hypothesized the "internal inducer" model that assuming all enzyme induction was caused by a generalized mechanism that was involved in both the synthesis of "inducible" enzymes (i.e., enzymes that were made only in the present of inducers) and the synthesis of "constitutive" enzymes (i.e., enzymes that were made at all times, no matter whether inducers were present or not). In inducible system, the normal *lacI* gene (termed *lacI*+) would produce the inducible enzyme to *inactivate* the internal or endogenous inducer. As a result, the system required an external inducer to activate the *lacZ* and *lacY* genes. In the constitutive system, the mutant *lacI* gene termed (termed *lacI*-) would produce the constitutive enzymes to deactivate the *lacI*+ gene so that allowed the synthesis of the internal inducers. As a

result, no external inducer was needed to activate the *lacZ* and *lacY* genes (Pardee, Jacob, and Monod 1959:174). In addition, they also hypothesized that when the *lacI+* and *lacI-* genes were both present in the cell at the same time, the constitutive system would dominate over the inducible system because of the *lacI-* gene had some stimulating effect on enzyme production.

For these assumptions, the zygote with the *lacZ+* gene and the *lacI+* gene would not produce β-galactosidase unless the external inducer was added because of the *lacI+* gene; the zygote with the *lacZ+* and *the lacI-* would produce β-galactosidase at all times, no matter whether the inducer was present or not, because of the the *lacI-* gene; the zygotes with the *lacZ-* and the *lacI+* genes and the *lacZ-* and the *lacI-* genes wouldn't produce β-galactosidase under any condition because of the *lacZ-* gene. In order to test the internal inducer model, especially the existence of the assuming internal inducers of the constitutive system, the scientists isolated various kinds of normal genes and mutant genes. One of the normal kinds was the *lacZ+* gene that could produce β-galactosidase when met with the *lacI+* and the other normal kind was the *lacI+* gene that could inactivate the internal inducer to keep the system being *inducible* state. One of the mutant kinds was the *lacZ-* gene that lost the capacity to produce β-galactosidase and the other kind was the *lacI-* gene that could produce β-galactosidase constitutively.

5.3 Unexpected phenomena discovered by ensuing experiments

In the stage of testing the internal inducer model, the scientists arranged male bacteria (donor) containing mutant *lacI-* and *lacZ-* genes to mate with female bacteria (recipient) containing normal *lacI+* and *lacZ+* genes, in the absence of inducers. As mentioned above, the scientists predicted that the enzyme synthesis would produce β-galactosidase. In internal inducer model's term, the system would converse from inducible state to constitutive state because of the dominance effect from the male's *lacI-* gene. But to their surprise, the synthesis system did not work. The scientists reasoned that the inducible *lacI+* allele should be dominant over the constitutive *lacI-* allele. The unexpected phenomenon became an anomaly to be investigated. No new model can guide scientist to design new experiments.

In such a situation, the scientists tried to understand the anomaly with the classical genetic reasoning. "This suggests that the dominant allele is the inducible (*i+*). If so, the *i+* should eventually become expressed in mating of type (B) ─ i.e., the zygotes, initially constitutive, should eventually become inducible." (Pardee, Jacob, and Monod 1959:174) In order to investigate the unexpected phenomenon, the scientists performed the experiments in opposite matting direction, that was, arranged male bacterium containing normal *lacI+* and *lacZ+* genes to mate with female bacterium containing mutant *lacI-* and *lacZ-* genes, both in the absent and in the present of inducer. As mentioned above, the scientists predicted that the enzyme synthesis would begin to produce β-galactosidase constitutively. Because when male's *lacZ+* genes entered females' cell, where the mutant *lacI-* gene would produce the assumed internal inducer, the zygote should begin synthesize β-galactosidase and without stopping. But to their surprise again, the synthesis began but *stopped* after about two hours (see the horizontal line in Fig. 3)

Why would the β-galactosidase synthesis interrupt in the circumstances of the presence of assumed internal inducer? Just then, when male's *lacI*+ genes subsequently entered female's cell, the scientists added external inducers, they found that the synthesis was resumed (see the upper and right curve in Fig. 3). How did *lacI*+ genes affect the synthesis to work again? The scientists reasoned, "When inducer added at this stage, enzyme synthesis is resumed, showing that the initially constitutive *z*+*i*+/*z*-*i*- zygotes have not been inactivated, but have become inducible." (Pardee, Jacob, and Monod 1959:175) Such the experimental outcome and explanation were considered to falsify the internal inducer model.

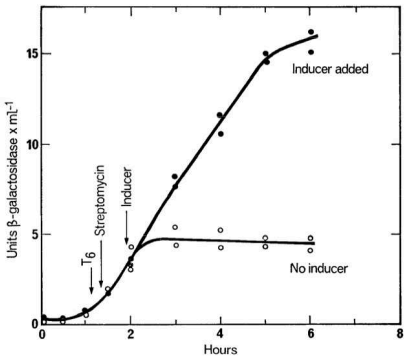


Figure 3. What the scientists observed was that after conjugation, the synthesis began in the absence of inducers, but only for a few hours. After those few hours, the synthesis resumed only in the adding of inducers externally. Reproduced from Pardee, Jacob, and Monod (1959). p. 173.

For a long time, Monod had not completely abandoned the internal inducer model , even attended a seminar which given by Leo Szilard. However, after enough consideration, Monod had started to accept the idea that a kind of the positive activities (induction) could be caused by the effect of double negative activities (the repression of repressors). Then they proposed "the repressor model". The *lacI*+ genes produced the "repressors" (later known were proteins). When no lactose was present in the environment, the repressors would block the synthesis by binding to the operator. When only lactose was present in the environment, the repressors would be inhibited by lactose, thus the genes would be able to synthesize β-galactosidase.

> According to the other, or 'repressor', model the activity of the galactosidase-forming system is inhibited in the wild typify a specific 'repressor' synthesized under the control of the *i*+ gene. The inducer is required only in the wild-type as an antagonist of the repressor. In the constitutive (i-), the repressor is not formed, or is inactive, hence the requirement for an inducer disappears." (Pardee, Jacob, and Monod 1959:176)

Had several new experimental evidences, the scientists had confirmed that the existence of the repressors and convinced that the repression effects had justify the repressor model. The repressor model was not only more empirically adequate than the in-

ternal inducer model but also more simpler than it. Because the repressor model did not require an abstract and nonexistent entity.

5.4 Multiple interventions for different aims and with different modes

The PaJaMo experiment was really a series of interventional experiments for one and the same subject. The experimentation involved multiple interventions for different aims and with different modes. It started with testing a given hypothesis, discovered unexpected phenomena, and then turned to toward an explorative aim without direction of any given hypothesis. After a series of investigation, the scientists provided a new model to solve and explain the novel phenomenon.

In the stage of testing the internal inducer model, the scientists produced the required mutants, to put them in the prepared experimental system, and to entice them to mate with one other. They intervened male bacterium containing the *lacI-* and *lacZ-* genes to mate with female bacterium containing the *lacI+* and *lacZ+*. The experimental result was no synthesis occurred. This is the case of the mode of *vertical* intervention with *inhibitory* effect. In the stage of exploring the new phenomenon, the scientists intervened male bacterium containing the *lacI+* and *lacZ+* genes to inject into female's cell containing the *lacI-* and *lacZ-*. The experimental result was that the synthesis began but stopped later. The immediate synthesis is the result of the mode of *vertical* intervention with *excitatory* effect (because of the *lacZ+* genes). Once the *lacI+* genes also entered the cell, the genes started to produce the repressors. When enough time had passed, the synthesis stopped. This is the result of the mode of *vertical* intervention with *inhibitory* effect. Later, the scientists added the external inducers, then the synthesis re-continued. It is the mode of *horizontal* intervention with *excitatory* effect. Because the scientists engineered the upstream stage of the mechanism of transcription and observed the positive changes of the downstream stage of the mechanism.

The PaJaMo experiment shows us that how multiple interventional modes used in a series of experiments. It also shows us that how interventional experiments used as essential means to realize plural experimental aims.[6] As Morange said, "it (the PaJaMo experiment) represented the final step in the development of a new vision of biology that had been begun in the 1940s and that made the question of the information contained in genes an ordering principle of all life." (1998:159) It brings about the discovery of the mechanism of gene regulation and leads molecular biology entering into a new stage.

6. Categories of Experimentation and the Nature of the PaJaMo Experiment

---

[6] With regarding to the investigation of new phenomenon is one of the important experimental aims, we may think of Water's "genetic approach" (Water 2004, 2008). Actually, in our view, the PaJamo experiment generally fits the genetic approach. The scientists *artificially produced mutants* (used required mutants), *gave genetic analyses* (as the scientists said "The suggests that the dominant allele is the inducible (*i+*)…From these observation we may conclude that the constitutive (*i-*) allele is inactive…" (Pardee, Jacob, and Monod 1959:174-175)), and *recombined the mutant to reveal a new biological process* (as the scientists said "…this is precisely the case…is a very strong argument in favor of the repressor model." (Pardee, Jacob, and Monod 1959:174-175)). All these steps leads to the scientific advancement. But Waters did not address the part of the experimental intervention in details.

For a long period of time, people believe that experimentation has only a single aim and function, i.e., testing theories. Since the end of the 20th century, the exploratory aim and function of experimentation has been revealed by some philosophers of science (Burian 1997; Steinle 1997) and gotten much attention (Burian 2007; Elliot 2007; Franklin 2005; O'Malley 2007; Steinle 2002; Waters 2004, 2007). Steinle (1997, 2002)'s distinction between theory-driven experimentation and exploratory experimentation has become a standard frame of categorizing experimentation (Franklin 2005: 888-889; O'Malley 2007: 339; Burian 2007: 286-288). However, these philosophers also emphasize that the distinction between theory-driven and exploratory experimentation does not mark a sharp division and the latter is not free of theory (See Waters 2007: 277-279).[7] Waters further notes the difference between being theory-directed and being theory-informed and the distinction between exploratory and theory-driven experimentation is made by the ways in which an experiment depends on theory (2007: 277). Nevertheless, these philosophers agree the two distinctive categories of experimentation, although not sharp, largely works for methodological analyses. However, we wonder whether or not there would be an experiment or a series of organized experiments which was used to both test hypotheses and explore novel things. If there is one, then what category we should classify it into? The experiment might test a hypothesis, falsify it, find anomalous phenomena, and then enter into an unknown field and become exploratory. Thus, we should say that the experiment is both theory-driven and exploratory. According to the previous discussion, we think that the PaJaMo experiment discussed in the previous section is the just one. As a consequence, this issues a challenge to the two basic categories of theory-driven and exploratory experimentation.

Elliot develops a taxonomy of exploratory experiments by discerning different kinds according to the three relatively independent dimensions: aims of experimental activity, role of theory in the activity, and methods or strategies for varying parameters (2007: 324). According to his taxonomy, "testing a hypothesis" is neither an aim of nor plays a role in an exploratory experiment. The aims of exploratory experiments include (1) "identifying regularities and developing new concepts," (2) "isolating or manipulating particular entities or phenomena," (3) "developing experimental techniques, instrumentation, or simulations," and (4) "resolving anomalies." The PaJaMo experiment explicitly realized aim 1 and 4. In order to resolve the classifying problem, one may simply add a third hybrid category of experimentation to the dichotomous categories. As a consequence, the PaJaMo experiment should be classified to

[7] All Burian (2007), Elliot (2007), and O'Malley (2007) emphasize this point. Burian points that there are a sharp methodological division between advocates of "hypothesis driven science" and advocates of "data-driven science". However, he emphasizes "[I]t is important to reduce the sharpness of this supposed dichotomy." (2007: 286-287) Elliot (2007) claims that theory plays a minimal role relative to other forms of experimentation and characterizes a few roles of theory in exploratory experimentation (2007: 324). O'Malley examines the interaction between exploratory and theory-driven experimentation within the context of an exploratory program of research.

the third hybrid category.[8] However, this resolution brings the new problem of the proliferation of varieties of hybrid experiments: for examples, experiments for both aim of testing and aim of manipulating particular entities, experiments for both aim of testing and aim of developing techniques, and so on. The point is to categorize experimentation according to some relatively independent dimensions or criteria as those Elliot (2007) has provided.

At this point, let us discuss Elliot's three independent dimensions. First, we can simply add the aim of testing hypotheses, model, and theories to the dimension of "aims of experimental activities." Second, we take Waters' distinction between "theory-directed" and "theory-informed" as two sub-dimensions of "role of theory in experimentation." In order to offer a more full taxonomy of the second dimension, we want to add a third sub-dimension "theory-free".[9] As for Elliot's third dimension of "methods or strategies for varying parameters, we want to introduce the two sub-dimensions of "interventional" and "non-interventional" to match up the topic of this paper, in which we draw a taxonomy of interventional modes from Craver and Darden's taxonomy of experiments. To sum up the previous discussion, we build the following table:

Dimensions of experimentation	Varying Characteristics of Experimentation with the Dimensions
Aims of Experimental Activities	Testing hypotheses Identifying regularities Developing new concepts Isolating or manipulating entities Developing techniques and instrumentation Simulating Resolving Anomalies
Role of Theory in Experimentation	Theory-directed: Testing hypotheses; Identifying regularities; Developing new concepts, etc. Theory-informed: Providing background information; Serving as a starting point or foil; New theory being constituted by exploratory projects, etc. Theory-free

[8] The other resolution is the appealing to Waters' distinction between experiments and programs of investigative research which combines explanatory reasoning with investigatory strategies (Waters 2004, 2007, 2008). According to this resolution, the PaJaMo experiment should be treated as a part of a program of investigative research for the regulatory mechanism of genes. However, our aim in this paper is only to analyze the PaJaMo experiment.

[9] Whether or not there are theory-free experiments depends on our interpretation or theory of the term "theory". Here we take a narrower interpretation of "theory" and set up the sub-dimension "theory-free".

Methods or Strategies of Varying Parameters	Interventional: [1] Interventional direction: Bottom-up or Top-down or Inter-stage [2] Interventional effect: Excitatory or Inhibitory Non-interventional

Table 1: A taxonomy of characteristics of experimentation according to three relatively dimensions

We believe that we can provide a more complete analysis of interventional experiments in molecular biology according to the framework of three experimental dimensions presented in the Table 1.

7. Concluding remarks

We have argued that multiple experimental interventions are frequently used in biological practice. We have provided a taxonomy of modes of experimental interventions that are developed from Craver and Darden's taxonomy of experiments. The taxonomy of interventional modes is built according the three directions (top-down, bottom-up, and inter-stage) and two effects (excitatory and inhibitory). We reexamine the famous PaJaMo experiment (a series of experiments for a single subject) to illustrate the new taxonomic framework. We find that scientists pursue both aims of testing hypotheses and investigating new phenomena. This motivate us to further consider the possibility that a series of experiments is performed to realize different aims of experimentation by using different strategies or methods. As a result, we provide a new taxonomy of characteristics of experimentation in which the molecular biological practice is adequately analyzed in the light of multiple aims and interventions.

References

Bechtel, William and Adele Abrahamsen (2005). "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421-441.

Burian, Richard M. (1997). "Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938-1952." *History and Philosophy of the Life Science* 19:27-45.

Burian, Richard M. (2007). "On microRNA and the Need for Exploratory Experimentation in Post-Genomic Molecular Biology." *History and Philosophy of the Life Sciences* 26:285-311.

Chen, Ruey-Lin (2013). "Experimental Discovery, Data Models, and Mechanisms in Biology: An Example from Mendel's Work." In Chao, Hsiang-Ke, Szu-Ting Chen, and Roberta L. Millstein (eds.). *Mechanism and Causality in Biology and Economics*. Dordrecht: Springer Press.

Chen, Ruey-Lin (2017). "Mechanisms, Capacities, and Nomological Machines: Integrating Cartwright's account of nomological machines and Machamer, Darden and Craver's account of mechanisms." In Chao, Hsiang-Ke, Szu-Ting Chen, and Julian Reiss (eds.). *Philosophy of Science in Practice: Nancy Cartwright and the Nature of Scientific Reasoning*. New York: Springer Press.

Cohen, G. N. and Jacques Monod (1957). "Bacterial Permeases." In Andre Lwoff and Agnes Ullmann (eds.) *Selected Papers in Molecular Biology by Jacques Monod*. Academic Press Inc.

Craver, Carl F. and Lindley Darden (2001). "Discovering Mechanisms in Neurobiology: The case of Spatial Memory." In Machamer, Peter, R. Grush, and P. McLaughlin (eds.) *Theory and Method in the Neuroscience*. Pittsburgh: University of Pittsburgh Press. Reprinted in Darden 2006, ch. 2.

Craver, Carl F. (2002). "Interlevel Experiments, Multilevel Mechanisms in the Neuroscience of Memory." *Philosophy of Science* (Supplement) 69:S83-S97.

Craver, Carl F. and Lindley Darden (2005). "Introduction: Mechanisms Then and Now." In Craver, Carl F. and Lindley Darden (eds.) *Studies in History and Philosophy of Biological and Biomedical Science*. Special Issue, "Mechanisms in Biology" 36:233-244.

Craver, Carl F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.

Craver, Carl F. (2008). "Axelrod, Julius" In Noretta Koertge (ed.) *New Dictionary of Scientific Biography*. Detroit, MI: Charles Scribner's Sons/Thomson Gale.

Craver, Carl F. and Lindley Darden (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. Chicago: The University of Chicago Press.

Darden, Lindley (1991). *Theory Change in Science: Strategies from Mendelian Genetics*. Oxford: Oxford University Press.

Darden, Lindley (2006). *Reasoning in Biological Discoveries: Essay on Mechanisms, Interfield Relations, and Anomaly Resolution*. Cambridge: Cambridge University Press.

Darden, Lindley (2013). "Mechanisms versus Causes in Biology and Medicine." In Chao, Hsiang-Ke, Szu-Ting Chen, and Roberta L. Millstein (eds.). *Mechanism and Causality in Biology and Economics*. Dordrecht: Springer Press.

Darden, Lindley and Carl.F. Craver (2002). "Strategies in the Interfield Discovery of the Mechanism of Protein Synthesis." *Studies in History and Philosophy of Biological and Biomedical Sciences* 33:1-28. Corrected and reprinted in Darden 2006, ch. 3.

Elliott, Kevin C. (2007). "Varieties of Exploratory Experimentation in Nanotoxicology." *History and Philosophy of the Life Sciences* 29 (3): 313-336.

Fangan, Melinda B. (2016). "Interventionist Omissions: A Critical Case Study of Mechanistic Explanation in Biology." *Philosophy of Science* 83:1082-1097.
Glennan, Stuart S. (1996). "Mechanisms and the Nature of Causation." *Erkenntnis* 44:49-71.

Franklin, Laura R. (2005). "Exploratory Experiments." *Philosophy of Science 72*: 888-899.

Glennan, Stuart S. (2002). "Rethinking Mechanistic Explanation." *Philosophy of Science (Supplement)* 69:S342-S353.

Hogness, David S., Melvin Cohn and Jacques Monod (1955). "Studies on the Induced Synthesis of β-galactosidase in Escherichia Coli: The Kinetics and Mechanims of Sulfur Incorporation." In Andre Lwoff and Agnes Ullmann (eds.) *Selected Papers in Molecular Biology by Jacques Monod*. Academic Press Inc.

Jacob, Francois and Jacqes Monod (1961). "Genetic Regulatory Mechanisms in the Synthesis of Protein." *Journal of Molecular Biology* 3:318-356.

Judson, Horace F. (1996). *The Eighth Day of Creation: The Makers of the Revolution in Biology*. Expanded Edition. Cold Spring Harbor., NY: Cold Spring Harbor Laboratory Press.

Machamer, Peter, Lindley Darden, and Carl F. Craver (2000). "Thinking About Mechanisms." *Philosophy of Science* 67:1-25.

Morange, Michel (1998). *A History of Molecular Biology*. Cambridge, Mass: Havard University Press.

Monod, Jacques (1947). "The Phenomenon of Enzymatic and its adaptation: And its Bearing on Problem of Genetics and Cellular Differentiation." In Andre Lwoff and Agnes Ullmann (eds.) *Selected Papers in Molecular Biology by Jacques Monod*. Academic Press Inc.

Monod, Jacques (1950). "Adaptation, Mutation and Segreation in the Formation of Bacterial Enzymes." In Andre Lwoff and Agnes Ullmann (eds.) *Selected Papers in Molecular Biology by Jacques Monod*. Academic Press Inc.

Monod, Jacques (1956). "Remarks on the Mechanism of Enzyme Induction." In Andre Lwoff and Agnes Ullmann (eds.) *Selected Papers in Molecular Biology by Jacques Monod*. Academic Press Inc.

Monod, Jacques (1958). "An Outline of Enzyme Induction." *Recueil* 77(6):569-585.

Monod, Jacques ([1965] 1977). "From Enzymatic Adaptation to Allosteric Transition." Reprinted in *Nobel Lectures in Molecular Biology: 1933-1975*, pp. 259-82. New York: Elsevier.

O'Malley, Maureen (2007). "Exploratory Experimentation and Scientific Practice: Metagenomics and the Proteorhodopsin Case." *History and Philosophy of Life Science* 29:335-358.

Pardee, Arthur B., François Jacob and Jacques Monod (1958). "Sur l'expression et le rôle des allèles «inductible» et «constitutif» dans la synthèse de la β-galactosidase chez des zygotes d'Escherichia Coli." [The role of the inducible alleles and the constitutive alleles in the synthesis of beta-galactosidase in zygotes of Escherichia cloi.] C. R. Hebd Seances Acad. Sci. 246:3125-8.

Pardee, Arthur B., François Jacob and Jacques Monod (1959). "The Genetic Control and Cytoplasmic Expression of 'Inducibility' in the Synthesis of β-galactosidase by E. Coli." In Andre Lwoff and Agnes Ullmann (eds.) *Selected Papers in Molecular Biology by Jacques Monod*. Academic Press Inc.

Pardee, Arthur (1979). "The PaJaMa Experiment." In Lwoff, Andre and Agnes Ullmann (eds.) *Origins of Molecular Biology: A Tribute to Jacque Monod*, pp. 109-116. New York: Academic Press.

Popper, Karl, R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson Edu cation.

Schaffner, Kenneth (1974a). "Logical of Discovery and Justification in Regulatory Genetics." *Studies in History and Philosophy of Science* 4:349-385.

Schaffner, Kenneth (1974b). "The Peripherality of Reductionism in the Development of Molecular Biology." *Journal for the History of Biology* 7:111-139.
Schaffner, Kenneth (1993). *Discovery and Explanation in Biology and Medicine*. Chicago, IL: University of Chicago Press.
Steinle, Friedrich (1997). "Entering New Fields: Exploratory Uses of Experimentation." *Philosophy of Science* (Proceedings) 64:S65-S74.
Steinle, Friedrich (2002). "Experiments in History and Philosophy of Scence." *Per spectives on Science* 10: 408-432.
Waters, C. Kenneth (2004). "What was Classical Genetics?" *Studies in History and Philosophy of Science* 35:783-809.
Waters, C. Kenneth (2007). "Causes that Make a Difference." *Journal of Philosophy* 104 (11): 551-579.
Waters, C. Kenneth (2008). "Beyond Theoretical Reduction and Layer-Cake Antireduction: How DNA Retooled Genetics and Transformed Biological Practice." In Michael Ruse (ed.) *The Oxford Handbook of Philosophy of Biology*. Oxford: Oxford University Press.
Weber, Marcel (2005). *Philosophy of Experimental Biology*. New York: Cambridge University Press.
Woodward, James (2002). "What Is a Mechanism? A Counterfactual Account." *Philosophy of Science (Supplement)* 69:S366-S377.
Woodward, James (2003). *Making Thing Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Woodward, James (2006). "Sensitive and Insensitive Causation." *Philosophical Review* 115:1-50.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association

Seattle, WA; 1-4 November 2018

Version: 31 October 2018

PhilSci ARCHIVE

Table of Contents

Page

Mixed-Effects Modeling and Non-Reductive Explanation

(4975 words)

Abstract: This essay considers a mixed-effects modeling practice and its implications for the philosophical debate surrounding reductive explanation. Mixed-effects modeling is a species of the multilevel modeling practice, where a single model incorporates simultaneously two (or even more) levels of explanatory variables to explain a phenomenon of interest. I argue that this practice makes the position of explanatory reductionism held by many philosophers untenable, because it violates two central tenets of explanatory reductionism: single level preference and lower-level obsession.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -2-

1. Introduction

Explanatory reductionism is the position which holds that, given a relatively higher-level phenomenon (or state, event, process, etc.), it can be reductively explained by a relatively lower-level feature (Kaiser 2015, 97; see also Sarkar 1998; Weber 2005; Rosenberg 2006; Waters 2008).[1] Though philosophers tend to have slightly different conceptions of the position, two central tenets of the position can still be extracted:[2]

> **Single level preference:** a phenomenon of interest can be fully explained by invoking features that reside at a single, well-defined level of analysis (e.g., molecular level in biology).

[1] According to Sarkar (1998), explanatory reduction is an epistemological thesis which is distinguished from constitutive (ontological) and theory reductionism theses. Kaiser further distinguishes two sub-types of explanatory reduction: (a) "a relation between a higher-level explanation and a lower-level explanation of the same phenomenon" (2015, 97); (b) individual explanations, i.e., given a relatively higher-level phenomenon, it can be reductively explained by a relatively lower-level feature (*Ibid*., 97). This essay will focus on the second sub-type. Besides, when referring to levels I mean either hierarchical organization such as universities, faculties, departments etc., or functional organization such as organs, tissues, cells etc. When referring to scales I mean spatial or temporal scaling where levels are not so clearly delimited.

[2] Similar summary of the position can be found in Sober (1999).

Seattle, WA; 1-4 November 2018 -3-

> **Lower-level obsession:** lower-level features always provide the most significant and detailed explanation of the phenomenon in question, so a lower-level explanation is always better than a higher-level explanation.

Philosophers sometimes express these two tenets explicitly in their work. For example, Alex Rosenberg holds that "[…] there is a full and complete explanation of every biological fact, state, event, process, trend, or generalization, and that this explanation will cite only the interaction of macromolecules to provide this explanation" (Rosenberg 2006, 12). Marcel Weber expresses a similar idea in his explanatory hegemony thesis, according to which it's always some lower-level physicochemical laws (or principles) that ultimately do the explanatory work in experimental biology (Weber 2005, 18-50). John Bickle attempts to motivate a 'ruthless' reduction of psychological phenomena (e.g., memory) to the molecular level (Bickle 2003).

However, many philosophers have questioned the plausibility of the position on the basis of scientific practice (Hull 1972; Craver 2007; Bechtel 2010; Brigandt 2010; Hüttemann and Love 2011; Kaiser 2015). To counter that position, some authors have pointed to the relevance of an important practice that has not received sufficient attention before: multiscale or multilevel modeling or sometimes called integrative modeling approach, where a set of distinct models ranging over multiple levels or scales—including the macro-phenomenon level/scale—are involved in explaining a (often complex) phenomenon of interest

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -4-

(Mitchell 2003, 2009; Craver 2007; Brigandt 2010, 2013a, 2013b; Knuuttila 2011; Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017). Often these models work together by providing diverse constraints on the potential space of representation (Knuuttila and Loettgers 2010; Knuuttila 2011; Green 2013).

This multilevel modeling surely casts some doubt on explanatory reductionism, for it seems unclear what reductively explains what—all those facts in the set of models ranging over different levels/scales are involved in doing some explanatory work. However, there is a species of multilevel modeling that has slipped away from most philosophers' sights: mixed-effects modeling (MEM hereafter)—also called multilevel regression modeling, hierarchical linear modeling, etc.—in which a single model incorporating simultaneously two (or even more) levels of variables is used to explain a phenomenon. For a mixed-effects model to explain, features of the so-called reducing and reduced levels must be simultaneously incorporated into the model, that is, they must go hand in hand.

MEM deserves special attention because it sheds new light on the reductionism-antireductionism debate by showing that (a) a mixed-effects model violating the two central tenets of explanatory reductionism can provide successful explanation, and (b) a single mixed-effects model without integrating with other epistemic means can also provide such successful explanation. Therefore, MEM first further challenges the explanatory reductionist position, and

second offers a novel perspective bolstering the multilevel/multiscale integrative approach discussed by many philosophers.

The essay proceeds as follows. Section 2 discusses the challenges faced by the traditional single-level modeling approach, and examines the reasons why the MEM approach is preferable in dealing with these challenges. Section 3 describes a MEM practice using a concrete model. Section 4 elaborates on the implications of MEM for the explanatory reductionism debate. Finally, Section 5 considers potential objections to my viewpoint.

2. Challenges to Reductive Explanatory Strategies

In many fields (e.g., biological, social and behavioral sciences) scientists find that the data collected show an intrinsically hierarchical or nested feature. Consider a simple example: we might be interested in examining relationships between students' achievement at school (A hereafter) and the time they invest in studying (T).[3] In conducting such a research, we might collect data from different classes (say 5 classes in total), with each class providing the same number of samples (say 10 students in each class). The data collected among classes might be taken for granted to be independent. Then we may use certain traditional statistical techniques such as ordinary least-squares (OLS) to analyze the data and build a linear relationship between A and T.

[3] For scientific studies of this kind, see Schagen (1990), Wang and Hsieh (2012), and Maxwell et al. (2017).

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -6-

However, this single-level reductive analysis can lead to misleading results, because it ignores the possibility that students within a class may be more similar to each other in important aspects than students from different classes. In other words, each group (class) may have its own features relevant to the relationship between A and T that the other groups lack. Hence, the data collected from the students are in fact not independent, i.e., the subjects are not randomly sampled, because the individuals (students) are clustered within groups (classes). In technical terms, we say our analysis may fall prey to the *atomistic fallacy* where we base our analysis solely on the individual level—i.e., we reduce all the group-level features to the individuals. Therefore, traditional OLS techniques such as multiple regression cannot be employed in this context, because the case under consideration violates a fundamental assumption of these techniques: the independence of observations (Nezlek 2008, 843).

Conversely, we may face the same problem the other way around if we fail to consider the inherently nested nature of the data. Consider the student-achievement-at-school case again. We may observe that in classes where the time of study invested by students is very high, the achievements of the students are also very high. Given such an observation, we may reason that students who invest a lot of time in studying would be more likely to get higher achievements at school. However, this inference commits the *ecological fallacy*, because it attributes the relationship observed at the group-level to the individual-level (Freedman 1999). The individuals may exhibit within-group differences that the single group-level analysis fails to capture. In technical terms, this inference flaws

because it reduces the variability in achievement at the individual-level to a group-level variable, and the subsequent analysis is solely based on group's mean achievement results (Heck and Thomas 2015, 3). Again, traditional statistical techniques such as multiple regression cannot be employed in this context.

In sum, a single-level modeling approach that disrespects the multilevel data structure can commit either an atomistic or an ecological fallacy. Confronted with these problems, one response is to 'tailor' the traditional statistical techniques by, e.g., adding an effect variable to the model which indicates the grouping of the individuals. However, many have argued that this approach is unpromising because it may give rise to enormous new problems (Luke 2004; Nezlek 2008; Heck and Thomas 2015). Alternatively, scientists have developed a new framework that takes the multilevel data structure into full consideration, i.e., the MEM approach, to which we now turn.

3. Case Study: A Mixed-Effects Model

Depending on different conceptual and methodological roots we have two broad categories of MEM approaches: the multilevel regression approach and the structural equation modeling approach. The former usually focuses on direct effects of predictor variables on (typically) a single dependent variable, while the latter usually involves latent variables defined by observed indicators (for details see Heck and Thomas 2015). For the purpose of this essay's arguments, I will concentrate on the first kind.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -8-

Consider the student-achievement-at-school example again. Since students are typically clustered in different classes, a student's achievement at school may be both influenced by her own features (e.g., time invested in studying) and her class's features (e.g., size of the class). Hence here comes two levels of analysis: the individual-level (level-1) and the group-level (level-2), and individuals ($i=1,2,\ldots,N$) are clustered in level-2 groups ($j=1,2,\ldots,n$).[4] Now suppose that students' achievements at school are represented as scores they get in the exam. The effect of time invested in studying on scores can be described as follows:

$$Y_{ij}=\beta_{0j}+\beta_{1j}X_{ij}+\varepsilon_{ij} \qquad (1)$$

where Y_{ij} refers to the score of individual i in the jth group, β_{0j} is a level-1 intercept representing the mean of scores for the jth group, β_{1j} a level-1 slope (i.e., different effects of study time on scores) for the predictor variable X_{ij}, and the residual component (i.e., an error term) ε_{ij} the deviation of individual i's score from the level-2 mean in the jth group. Equation (1) looks like a multiple regression model; however, the subscript j reveals that there is a group-level incorporated in the model. It can also be seen from this equation that both the intercept β_{0j} and slope β_{1j} can vary across the level-2 units, that is, different groups can have different intercepts and slopes.

[4] Note that, for instructive purposes, our case involves only two levels; however, the MEM approach can in principle be extended to many more levels.

Seattle, WA; 1-4 November 2018 -9-

The most remarkable thing of MEM is that we treat both the intercept and slope at level-1 as dependent variables (i.e., outcomes) of level-2 predictor variables. So here we write the following equations expressing the relationships between the level-1 parameters and level-2 predictors:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad (2)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \quad (3)$$

where β_{0j} refers to the level-1 intercept in level-2 unit j, γ_{00} denotes the mean value of the level-1 intercept, controlling for the level-2 predictor W_j, γ_{01} the slope for the level-2 variable W_j, and u_{0j} the error (i.e., the random variability) for unit j. Also, β_{1j} refers to the level-1 slope in level-2 unit j, γ_{10} the mean value of the level-1 slope controlling for the level-2 predictor W_j, γ_{11} the effect of the level-2 predictor W_j, and u_{1j} the error for unit j.

Equations (2) and (3) have specific meanings and purposes. They express how the level-1 parameters, i.e., intercept or slope, are functions of level-2 predictors and variability. They aim to explain variations in the randomly varying intercepts or slopes by adding one (or more) group-level predictor to the model. These expressions are based on the idea that the group-level characteristics such as group size may impact the strength of the within-group effect of study time on

scores. This kind of effect is called a *cross-level interaction* for it involves the impact of variables at one level of a data hierarchy on relationships at another level. We will discuss this in detail in the next section.

Now we combine equations (1), (2) and (3) by substituting the level-2 parts of the model into the level-1 equation. We finally obtain the following equation:

$$Y_{ij}=\left[\gamma_{00}+\gamma_{10}X_{ij}+\gamma_{01}W_j+\gamma_{11}X_{ij}W_j\right]+\left[u_{1j}X_{ij}+u_{0j}+\varepsilon_{ij}\right] \qquad (4)$$

This equation can be simply understood that Y_{ij} is made up of two components: the fixed-effect part expressed by the first four terms and the random-effect part expressed by the last three terms. Note that the term $\gamma_{11}X_{ij}W_j$ denotes a cross-level interaction between level-1 and level-2 variables, which is defined as the impact of a level-2 variable on the relationship between a level-1 predictor and the outcome Y_{ij}. We have 7 parameters to estimate in (4), they are four fixed effects: intercept, within-group predictor, between-group predictor and cross-level interaction, two random effects: the randomly varying intercept and slope, and a level-1 residual.

Now a mixed-effects model has been built, and the next step is to estimate the parameters of the model. However, we will skip this step and turn to explore the philosophical implications of the modeling practice relevant to the explanatory reductionism debate.

4. Implications for the Explanatory Reductionism Debate

Seattle, WA; 1-4 November 2018 -11-

Looking closely into the MEM practice, we find that a couple of important philosophical implications for the explanatory reductionism debate can be drawn.

4.1. All levels are indispensable

The first, and most obvious, feature of MEM is that it routinely involves many levels of analysis in a single model, and all these levels are indispensable to the model in the sense that no level can be reduced to or replaced by the other levels. These levels consist of both the so-called reducing level in the reductionist's terminology, typically a lower-level that attempts to reduce another level, and the reduced level, typically a higher-level to be reduced by the reducing level. In our student-achievement-at-school case, for example, a reductionist may state that the group-level will be regarded as the reduced level whereas the student-level as the reducing level.

The indispensability of each level in the model can be understood in two related ways. First, due to the nested nature of data, only when we incorporate different levels of analyses to the model can we avoid either the atomistic or ecological fallacy discussed in Section 2. As discussed in the student-achievement-at-school example where students are clustered in different classes (in the manner that students from the same class may be more similar to each other in important aspects than students from different classes), reducing all the analyses to the level of individual students can simply miss the important

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -12-

information associated with group-level features and thus lead to misleading results. Although it's true that the problem might be partially mitigated by tailoring traditional single-level analytical techniques such as multiple regression, it's also true that this somewhat ad hoc maneuver can simply bring about various new vexing and recalcitrant issues (Luke 2004; Nezlek 2008; Heck and Thomas 2015).

Second, the problem can also be viewed from the perspective of identifying explanatory variables. In building a mixed-effects model, the main consideration is often to find a couple of variables that may play the role of explaining the pattern or phenomenon observed in the data. Here a modeler must be clear about how to assign explanatory variables, for instance, she must consider if there are different levels of analyses and, if so, which explanatory variables should be assigned to what levels, and so on. These considerations may come before her model building because of background knowledge, which paves the way for her to develop a conceptual framework for investigating the problem of interest. However, without such a clear and rigorous consideration of identifying and assigning multilevel explanatory variables, an analysis can flaw simply because it confounds variables at different levels.

Respecting the multilevel nature of explanatory variables has another advantage: "Through examining the variation in outcomes that exists at different levels of the data hierarchy, we can develop more refined theories about how explanatory variables at each level contribute to variation in outcomes" (Heck and Thomas 2015, 33). In other words, in respecting the multilevel nature of

Seattle, WA; 1-4 November 2018 -13-

explanatory variables, we get a clear idea of how, and to what degrees, explanatory variables at different levels contribute to variation in outcomes. If these variables do contribute to variation in outcomes, as it always happens in MEM, then the situation suggests an image of *explanatory indispensability*: all the explanatory variables at different levels are indispensable to explaining the pattern or phenomenon of interest.

Given these considerations, therefore, one implication for the explanatory reductionism debate becomes clear: it isn't always the case that, given a relatively higher-level phenomenon it can be reductively explained by a relatively lower-level feature. Rather, in cases where the data show a nested structure or, put differently, the phenomenon suggests multilevel explanatory variables, we routinely combine the higher-level with the lower-level in a single (explanatory) model. As a result, one fundamental tenet of explanatory reductionism is violated: single level preference.

4.2. Interactions between levels

Another crucial feature of multilevel modeling is its emphasis on *a cross-level interaction*, which is defined as

> "The potential effects variables at one level of a data hierarchy have on relationships at another level […]. Hence, the presence of a cross-level interaction implies that the magnitude of a relationship observed within

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -14-

> groups is dependent on contextual or organizational features defined by higher-level units". (Heck and Thomas 2015, 42-43)

Remember that there is a term $\gamma_{11} X_{ij} W_j$ in our mixed-effects model discussed in Section 3, which indicates the cross-level interaction between the group-level and the individual-level. More specifically, this term can be best construed as the impact of a group-level variable, e.g., group size, upon the individual-level relationship between a predictor, e.g., study time, and the outcome, e.g., students' scores.

The cross-level interaction points to the plain fact that an organization or a system can somehow influence its members or components by constraining how they behave within the organization or system. This doesn't necessarily imply top-down causation (Section 5.3 will turn back to this point). Within the context of scientific explanation, however, it does imply that it isn't simply that characteristics at different levels separately contribute to variation in outcomes, but rather that they interact in producing variation in outcomes. In other words, the pattern or phenomenon to be explained can be understood as generated by the interaction between explanatory variables at different levels. Therefore, to properly explain the phenomenon of interest, we need not only have a clear idea of how to assign explanatory variables to different levels but also an unequivocal conception of whether these explanatory variables may interact.

Different models can be built depending on different considerations of the cross-level interaction. To see this, consider the student-achievement-at-school

Seattle, WA; 1-4 November 2018 -15-

example again. In some experiment setting we may assume that there was no cross-level interaction between group-level characteristics and the individual-level relationship (between study time and scores). In such a situation, we kept the effect of individual study time on scores the same across different classes, i.e., we kept the slope constant across classes. In the meanwhile, we treated another group-level variable (i.e., intercept) as varying across classes, i.e., different classes have different average scores. So, this is a case where we have a clear idea of how to assign explanatory variables but no consideration of the cross-level interaction. Nonetheless, in a different experiment setting we may assume that there existed cross-level interaction, and hence the effect of individual study time on scores can no longer be kept constant across different classes. At the same time, we treated another group-level variable (i.e., intercept) as varying across classes. Hence, this is a case where we have both a clear idea of how to assign explanatory variables and a consideration of the cross-level interaction. Corresponding to these two different scenarios, two different mixed-effects models can be built, as shown below:

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -16-

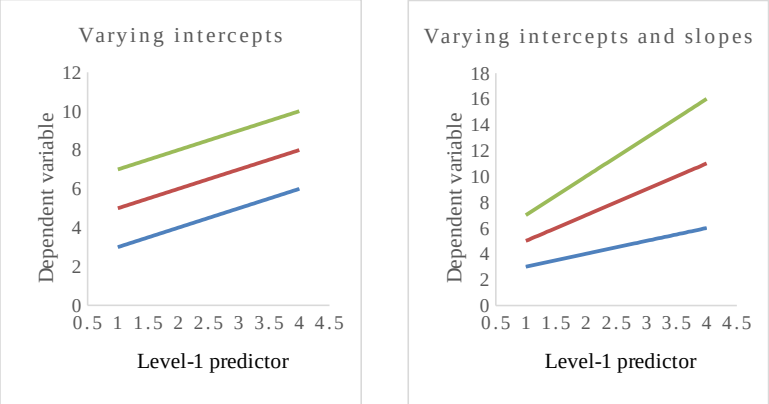


Figure 1. Two different models showing varying intercepts or varying intercepts and slopes, respectively. Three lines represent three classes. This figure is adapted from Luke (2004, 12).

Given such a cross-level interaction, therefore, the explanatory reductionist position has been further challenged. This is because any reductive explanation that privileges one level of analysis—usually the lower-level—over the others falls short of capturing this kind of interaction between levels. If they fail to do so, then they are missing important terms relevant to explaining the phenomenon of interest. As a consequence, a mixed-effects model involving interactions between levels simultaneously violates the two fundamental pillars of explanatory reductionism: first, it violates single level preference because it involves multilevel explanatory variables in explaining phenomena, and second, it violates lower-level obsession because it privileges no levels—all levels are interactively engaged in producing outcomes.

5. Potential Objections

This section considers two potential objections.

5.1. In-principle argument

One argument that resurfaces all the time in the reductionism-versus-antireductionism debate is the in-principle argument, the core of which is that even if reductive explanations in a field of study are not available for the time being, it doesn't follow that we won't obtain them someday (e.g., Sober 1999; Rosenberg 2006). Therefore, according to some reductionists, the gap between current-science and future-science is simply a matter of time, for advancement in techniques, experimentation and data collecting can surely fill in the gap.

However, I think the argument flaws. To begin with, advancement in techniques, experimentation and data collecting isn't always followed by reductive explanations. For example, in our MEM discussed in Section 3, even if the data about the individual-level is available and sufficiently detailed, it isn't the case that we explain the phenomenon of interest in terms of the data from the individual-level alone. Consider another example: in dealing with problems associated with complex systems in systems biology, even though large-scale experimentation (e.g., via computational simulation) can be conducted and high throughput data arranging over multiple scales/levels can be collected, a bottom-up reductive approach must be integrated with a top-down perspective so as to

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -18-

produce useful explanations or predictions (Green 2013; Green and Batterman 2017; Gross and Green 2017).

Nevertheless, reductionists may reply that the situations presented above only constitute an in-practice impediment, for it doesn't undermine the *possibility* that lower-level reductive explanations, typically provided by some form of 'final science', will be available someday. Let us dwell on the notion of possibility a bit longer. The possibility here may be construed as a *logical possibility* (Green and Batterman 2017, 21; see also Batterman 2017). Nonetheless, if it's merely logically possible that there will be some final science providing only reductive explanations, then nothing can exclude another logical possibility that there will be some 'mixed-science' providing only multilevel explanations. After all, how can we decide which logical possibility is more possible (or logically more possible)? I doubt that logic alone could provide anything useful in justifying which possibility is more possible, and that appealing to logical possibility could offer anything insightful in helping us understand how science proceeds. As Batterman puts, "Appeals to the possibility of *in principle* derivations rarely, if ever, come with even the slightest suggestion about how the derivations are supposed to go" (2017, 12; author's emphasis).

Another interpretation of possibility may be associated with real possibilities, referring to the actual cases of reductive explanations happening in science. Unfortunately, I don't think the real scenario in science speaks for the reductionist under this interpretation. Though it's impossible to calculate the absolute cases of non-reductive explanations occurring in science, a cursive look at scientific

Seattle, WA; 1-4 November 2018 -19-

practice can tell that a large portion of scientific explanations proceeds in a non-reductive fashion, as suggested by multilevel modeling (Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017; Mitchell and Gronenborn 2017). Moreover, even in areas such as physics which was regarded as a paradigm for the reductionist stance, progressive explanatory reduction doesn't always happen (Green and Batterman 2017; Batterman 2017).

In sum, we have shown that the in-principle argument fails for it neither offers help in understanding how science proceeds if it's construed as implying a logical possibility, nor goes in tune with scientific practice if it's construed as implying real possibilities.

5.2. Top-down causation

In Section 3 we have shown that there is a cross-level interaction taking the form that higher-level features may impact lower-level features. A worry arises: Does this imply top-down causation?

My answer to this question is twofold. First, it's clear that this short essay isn't aimed to engage in the philosophical debate about whether, and in what sense, there exists top-down causation (see Craver and Bechtel 2007; Kaiser 2015; Bechtel 2017). Second, what we can do now is to show that the cross-level interaction is a clear and well-defined concept in multilevel modeling. It unambiguously means the constraints on the lower-level processes exerted by the higher-level parameters (Green and Batterman 2017). In our multilevel modeling

discussed in Section 3, we have shown that group-level features may impact some individual-level features through the way that each group possesses its own feature relevant to explaining the differences at the individual-level across groups. This idea is incorporated into the mixed-effects model by assigning some explanatory variables to the group-level and a cross-level interaction term to the model.

The idea of cross-level-interaction-as-constraint is widely accepted in multilevel modeling broadly construed, where constraint is usually expressed in the form of initial and/or boundary conditions. For example, in modeling cardiac rhythms, due to "the influences of initial and boundary conditions on the solutions of the differential equations used to represent the lower level process" (Noble 2012, 55; Cf. Green and Batterman 2017, 32), a model cannot simply narrowly focus on the level of proteins and DNA but must also consider the levels of cell and tissue working as constraints. The same story happens in cancer research, where scientists are advocating the idea that tumor development can be better understood if we consider the varying constraints exerted by tissue (Nelson and Bissel 2006; Shawky and Davidson 2015; Cf. Green and Batterman 2017, 32).

6. conclusion

This essay has shown that no-reductive explanations involving many levels predominate in areas where the systems under consideration exhibit a hierarchical structure. These explanations violate the fundamental pillars of explanatory

Seattle, WA; 1-4 November 2018 -21-

reductionism: single level preference and lower-level obsession. Traditional single-level reductive approaches fall short of capturing systems of this kind because they face the challenges of committing either the atomistic or ecological fallacy.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -22-

References

Batterman, Robert. 2013. The "Tyranny of Scales." In *The Oxford Handbook of Philosophy of Physics*, ed. Robert Batterman, 255-286. Oxford: Oxford University Press.

———. 2017. "Autonomy of Theories: An Explanatory Problem." *Noûs* 1-16.

Bechtel, William. 2010. "The Downs and Ups of Mechanistic Research: Circadian Rhythm Research as an Exemplar." *Erkenntnis* 73:313–328.

———. 2017. "Explicating Top-Down Causation Using Networks and Dynamics." *Philosophy of Science* 84:253–274.

Bickle, John. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Brigandt, Ingo. 2010. "Beyond Reductionism and Pluralism: Toward an Epistemology of Explanatory Integration in Biology." *Erkenntnis* 73 (3): 295-311.

———. 2013a. "Explanation in Biology: Reduction, Pluralism, and Explanatory Aims." *Science and Education* 22:69–91.

———. 2013b. "Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:461–465.

Craver, Carl. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, Carl, and William Bechtel. 2007. “Top-down Causation without Top-Down Causes.” *Biology and Philosophy* 22:547–563.

Freedman, David. 1999. “Ecological Inference and the Ecological Fallacy.” In *International Encyclopedia of the Social and Behavioral* Sciences, vol. 6, ed. Neil Smelser, and Paul Baltes, 4027–4030. New York: Elsevier.

Green, Sara. 2013. “When One Model Isn’t Enough: Combining Epistemic Tools in Systems Biology.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:170–180.

Green, Sara, and Robert Batterman. 2017. “Biology Meets Physics: Reductionism and Multi-Scale Modeling of Morphogenesis.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 61:20–34.

Gross, Fridolin, and Sara Green. 2017. “The Sum of the Parts: Large-Scale Modeling in Systems Biology.” *Philosophy, Theory, and Practice in Biology* 9: (10).

Heck, Ronald, and Scott Thomas. 2015. *An Introduction to Multilevel Modeling Techniques* (3rd Edition). New York: Routledge.

Hull, David. 1972. “Reductionism in Genetics—Biology or Philosophy?” *Philosophy of Science* 39 (4): 491-499.

Hüttemann, Andreas, and Alan Love. 2011. “Aspects of Reductive Explanation in Biological Science: Intrinsicality, Fundamentality, and Temporality.” *British Journal for the Philosophy of Science* 62 (3): 519-549.

Kaiser, Marie. 2015. *Reductive Explanation in the Biological Sciences*. Springer.

Knuuttila, Tarja. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Studies in History and Philosophy of Science Part A* 42:262–271.

Luke, Douglas. 2004. *Multilevel Modeling*. London: SAGE Publications, Inc.

Maxwell, Sophie, Katherine Reynolds, Eunro Lee, et al. 2017. “The Impact of School Climate and School Identification on Academic Achievement: Multilevel Modeling with Student and Teacher Data.” *Frontiers in Psychology* 8:2069.

Mitchell, Sandra. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

———. 2009. *Unsimple Truths: Science, Complexity, and Policy*. Chicago: University of Chicago Press.

Nezlek, John. 2008. “An Introduction to Multilevel Modeling for Social and Personality Psychology.” *Social and Personality Psychology Compass* 2/2 (2008):842–860.

Noble, Daniel. 2012. “A Theory of Biological Relativity: No Privileged Level of Causation.” *Interface Focus* 2(1):55–64.

O’Malley Malley, Ingo Brigandt, Alan Love, et al. 2014. “Multilevel Research Strategies and Biological Systems.” *Philosophy of Science* 81:811–828.

Rosenberg, Alex. 2006. *Darwinian Reductionism, or How to Stop Worrying and Love Molecular Biology*. Chicago: University of Chicago Press.

Sarkar, Sahotra. 1998. *Genetics and Reductionism*. Cambridge: Cambridge University Press.

Schagen, I. P. 1990. “Analysis of the Effects of School Variables Using Multilevel Models.” *Educational Studies* 16:61–73.

Shawky, Joseph, and Lance Davidson. 2015. “Tissue Mechanics and Adhesion during Embryo Development.” *Developmental Biology* 401(1):152–164.

Sober, Elliot. 1999. “The Multiple Realizability Argument against Reductionism.” *Philosophy of science* 66:542–564.

Wang, Yau-De, and Hui-Hsien Hsieh. 2012. “Toward a Better Understanding of the Link Between Ethical Climate and Job Satisfaction: A Multilevel Analysis.” *Journal of Business Ethics* 105:535–545.

Waters, C. Kenneth. 2008. “Beyond Theoretical Reduction and Layer-Cake Antireduction: How DNA Retooled Genetics and Transformed Biological Practice”. In *The Oxford Handbook of Philosophy of Biology*, ed. Michael Ruse, 238-262. New York: Oxford University Press.

Weber, Marcel. 2005. *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

The Universe Never Had a Chance

C. D. McCoy[‡]

1 March 2018

Abstract

Demarest asserts that we have good evidence for the existence and nature of an initial chance event for the universe. I claim that we have no such evidence and no knowledge of its supposed nature. Against relevant comparison classes her initial chance account is no better, and in some ways worse, than its alternatives.
Word Count: 4712

1 Introduction

Although cosmology, the study of the universe's evolution, has largely become a province of physics, philosophical speculation concerning cosmogony, the study of the origin of the universe, continues up to the present. Certainly, many believe that science has settled this too by way of the well-known and well-confirmed big bang model of the universe. According to the big bang account the universe began in a extremely hot, dense state, composed of all the different manifestations of energy that we know. Indeed, time itself began with the big bang. Yet, properly speaking, the universe's past singularity is not some event in spacetime according to the general theory of relativity. In cosmological models this hot dense state called the big bang is generally understood instead as just a very early stage of the universe's evolution, i.e. properly a part of cosmology and not cosmogony. While we may be highly confident that the entire big bang story is correct back to a very early time, our confidence should at some point decrease as we near the supposed "first moment". Thus there remains world enough and time to engage in traditional philosophical and scientific speculations about cosmogony and cosmology alike. Were there previous stages to the universe? What brought the universe into existence? What was the character of this initial happening (should it in fact exist)?

The ubiquity of probabilities in modern physical theories, e.g. quantum mechanics and statistical mechanics, has led some to wonder as well how chance should fit into our

[*]**Acknowledgements:** Pending.

[†]School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK. email: casey.mccoy@ed.ac.uk

cosmogonical worldview. In this vein, Demarest (2016) argues that the probabilities of all events in a(n ostensibly) deterministic universe can be derived from an initial chance event and, what's more, that "we have good evidence of its existence and nature." In this paper I aim to dispute these latter claims. I argue that we do not have any evidence at all of an initial chance event in a big bang universe as described above, much less of its nature. What we rather have in Demarest's account is just a particular way of interpreting probabilistic theories, where all probabilities are taken to derive from ontic chances pertaining to the particular genesis of the relevant physical system, e.g. the universe as a whole. I claim that this interpretation, while coherent, should be disfavored in cosmology—we should rather say that *the universe never had a chance*.[1] Along the way I will make several clarifying remarks concerning the relation of chance and determinism, cosmological probabilities, and alternative interpretations of statistical and quantum mechanics.

2 Chance and Determinism in Physical Theory

By the *world* metaphysicians usually mean something like "the maximally inclusive entity whose parts are all the things that exist." Of course terminology varies. This particular rendering comes from Schaffer (2010, 33), who instead chooses to call this entity the *cosmos*. Cosmologists do not usually call their object of study the cosmos; more commonly they say that they study the *universe*. In *Cosmology: The Science of the Universe*, Harrison explicitly notes the philosophical and historical dimensions of the world taken in its broadest sense, designating this world as a whole the *Universe*. Cosmology, according to Harrison, is the study of universes, by which he means particular models of the Universe (Harrison, 2000, Ch. 1). Cosmological models are the particular concern of physical cosmologists; they are physical models of the Universe, which describe especially its large-scale structure and the evolution thereof.

In what follows I employ these terminologies in the following way. By the *world* I designate the locus of (principally) metaphysical questions concerning the Universe. Is the world deterministic? Is it chancy? By the *universe* I designate the locus of principally physical questions concerning the Universe. How did the big bang universe begin? How will it end? These are questions to which the big bang model should provide an answer.

I do not mean, of course, to introduce an admittedly arbitrary distinction between science and metaphysics by differentiating universes and worlds. Indeed, when one asks whether the world is deterministic, many metaphysicians of science would look first to models of the Universe to help decide the question. Wüthrich for example remarks, matter-of-factly, that "this metaphysical question deflates into the question of whether our best physical *theories* entail that the world is deterministic or indeterministic" (Wüthrich, 2011, 366).

[1] There are several senses, in fact, in which this claim is true. Cosmology suggests that the inevitable fate of the universe is to become ever more sparse and empty through the accelerated expansion of space under the influence of dark energy.

Indeed, many discussions of determinism adopt the approach mentioned by Wüthrich. Let *determinism* denote the thesis that the world is deterministic. Then, following for example (Lewis, 1983, 360), a world is *deterministic* if and only if the laws of that world are deterministic. To determine whether the laws of the universe are deterministic, we must look to our theories of which those laws are part and ask whether those laws taken together should be considered deterministic. It is by no means a straightforward matter to decide whether a given physical theory is deterministic of course. Even the classic example of deterministic physics, Newtonian mechanics, admits many counterexamples against its putative determinism (Earman, 1986; Norton, 2008). General relativity as well seemingly permits indeterministic phenomena in the form of causal pathologies (closed timelike curves) (Earman, 1995) and, if the hole argument is to be believed, is hopelessly rife with indeterminism (Earman and Norton, 1987).

Although classical theories like classical mechanics and general relativity are nevertheless debatably deterministic, surely probabilistic theories like quantum mechanics are properly characterized as indeterministic (at least so long as the probabilities involved are objective features of the world). Yet various interpretations of probabilistic theories seek to avoid indeterminism even here, where it seems unassailable, by characterizing probabilities as merely epistemic or subjective, or else by presenting them as fully deterministic theories (as in the Bohmian interpretation of quantum mechanics). Philosophers have raised serious concerns, however, over how one can truly understand probabilities in deterministic theories, an issue that has been termed the "paradox of deterministic probabilities" (Loewer, 2001; Winsberg, 2008; Lyon, 2011) in statistical mechanics, since objective probabilities seem to entail indeterminism necessarily.

The most well-known and successful reconciliation of chance and determinism in the context of statistical mechanics is defended by Loewer (2001). It is seldom recognized by interpreters, however, that there is no reconciliation in the sense of simultaneous compatibility between chance and determinism. The world cannot both be chancy and deterministic as a matter of metaphysical fact. As Lewis writes, "to the question of how chance can be reconciled with determinism, or to the question of how disparate chances can be reconciled with one another, my answer is: *it can't be done* (Lewis, 1986, 118). This is because chance entails indeterminism, the contrary of determinism. Thus, insofar as the probabilities of statistical mechanics and quantum mechanics are objective, these theories are indeterministic theories. Loewer's account actually shows us how deterministic laws can co-exist with indeterministic laws within a theory. The source of all probabilities in statistical mechanics, according to Loewer, is in an initial chance distribution over microscopic states of affairs. After the initial time these states of affairs evolve deterministically. Note that although for almost all times evolution is deterministic, it is not so at all times. There is an initial chance event, which is where the indeterminism of the theory appears. A deterministic theory is, recall, a theory whose laws are deterministic, not a theory whose laws are mostly deterministic or operate deterministically for almost all times.

Loewer's account is also presented in terms of Humean chances, so he does not believe

Seattle, WA; 1-4 November 2018 -29-

these chances and laws actually exist. According to the modern Humean, they merely are the result of the best systematizations of the occurrent facts, in keeping with Lewis's "best systems account" of laws and chances. Demarest, however, offers a small tweak to Loewer's Humean account by invoking a "robustly metaphysical account of chance" (Demarest, 2016, 256). She claims that such chances are compatible with determinism, and indeed they are when, as said, compatibility is understood to pertain to the co-existence of indeterministic and deterministic laws in a single theory—which, however, do not operate at the same time.[2]

Demarest's central claims are that this initial chance event exists and that we have good evidence for it. I dispute these claims in the remainder of the paper.

To begin, it is not so clear what exactly Demarest takes the evidence for the initial chance event to be. She does contrast the evidential position of her view with the Humean view of Loewer, claiming that, "for the Humean, the statistical patterns in the world are not evidence of an initial chance event" (Demarest, 2016, 261)—presumably this is so because Humeans reject the metaphysics of chance for the usual Humean reasons. One might suppose, then, that she believes that statistical patterns in the world are evidence of an initial chance event for all those who do not share the Humeans ontological worries. Let us accept, for the moment then, that statistical patterns may be *some* evidence for the existence of chances, for it is difficult to see what other evidence there might be for an initial chance event. In that case, on what grounds might we say that statistical patterns are good evidence for initial chances? I consider a series of three salient contrast classes.

First, do statistical patterns in data provide good evidence for indeterministic (i.e. chancy) theories *rather than deterministic theories*? It would seem that the answer is: not necessarily. (Werndl, 2009), for example, argues for the observational equivalence of indeterministic theories and deterministic theories. If one could contrive a fully deterministic theory that reproduces the same statistical patterns of the relevant phenomena observed in nature, then it would seem that such patterns provide no better evidence for the indeterministic theory than the deterministic one. However, since the theories under discussion, statistical mechanics and quantum mechanics, are generally characterized as indeterministic, let us flag but set aside the possibility of fully deterministic alternatives to them.

So, second, do statistical patterns provide good evidence for initial chances *rather than non-initial chances*? It would seem that the answer is firmly: no. There is a variety of ways one could implement chances into a probabilistic theory like statistical mechanics. All one must do, as Loewer shows us by example, is neatly separate when the indeterministic laws are operative and when the deterministic laws are operative. Loewer chooses to locate all the indeterminism in one place—the initial time—but one could equally locate it at another time, at many times, or even all times. Statistical mechanics does not wear its interpretation on its sleeve, just as quantum mechanics does not decide between solutions of the measurement problem, whether initial chances as in Bohmian mechanics or collapse

[2]Still, it is worth emphasizing that her claim that her account applies to deterministic worlds is false, for chancy worlds are not deterministic.

dynamics as in GRW (discrete time collapses) or CSL (continuous collapses). Unless there are evidential reasons to favor one implementation of indeteterministic probabilities over the others, there is not good evidence for an initial chance event. Certainly statistical patterns in nature will not do so.

Third, do statistical patterns provide good evidence for "robustly metaphysics" chances *rather than Humean chances*? It seems as if this might Demarest's intended contrast class, since much of the discussion in the paper concerns the Humean account. I will have something to say about the relative merits of Demarest's non-Humean account and Loewer's Humean account at the end of the next section. In any case though, it does not seem as if statistical patterns decide the matter in Demarest's mind, for she repeatedly demurs in the face of Humean responses to the considerations she raises, claiming only to offer an alternative "for philosophers who are antecedently sympathetic to governing laws of nature or powerful properties" (Demarest, 2016, 261-2). She finds it "plausible to think of the universe as having an initial state and as producing subsequent states in accordance with the laws of nature (some of which may be chancy)" (Demarest, 2016, 261). Such metaphysical intuitions are not grounded on observations of statistical patterns. Statistical patterns do not have any evidential bearing on the metaphysical dispute between the Humean and non-Humean.

Therefore, based on my canvassing of relevant alternatives, I conclude that we in fact do not have good evidence for an initial chance event, where evidence is interpreted in terms of statistical patterns (or in any usual sense of the term "evidence"). At best we have a motivation to attend to indeterministic theories when our evidence displays statistical patterns. It is another matter entirely to decide how to implement probabilities in that theory.

That said, Demarest's reasoning could be interpreted at points as invoking explanatory considerations as justification for the initial chance interpretation. Insofar as one considers "what justifies" as constituting evidence, perhaps these explanatory considerations should be counted as evidence.[3] Nevertheless, it does not look, on the face of it, like we have good evidence for an initial chance event still. Repeating the three cases considered before: deterministic and chancy theories can both serviceably explain statistical evidence; alternative implementations of chance in interpretations of indeterministic theories explain statistical evidence equally well; Humean and non-Humean metaphysics each render a story for how statistical patterns come about (merely subjective intuitions notwithstanding). Without explicit explanatory reasons to prefer one of these alternatives to the other, reasons lacking in Demarest's argument, good evidence (in this wider sense) for an initial chance event remains elusive.

[3]There are obvious dangers with going to far in this direction. Suppose that the Supreme Being explains all. Then it would appear that we have very good evidence of Its existence, which is obviously absurd.

3 Chance and Determinism in Systems of the World

In the previous section I gave reasons to doubt Demarest's claims about an initial chance event and our evidence for it. I disputed especially that we have evidence for it and did so by comparing it to alternatives of three different kinds. In the first case I characterized the issue (in part) as a matter of theory choice, namely of choosing between an indeterministic and deterministic theory. In the second case I characterized the issue as a matter of theory interpretation, namely of interpreting between different ways of implementing probability in a theory that does not decide one way or another on how this must be done. In the third case I characterized the issue as a matter of metaphysics, namely of deciding between the ontological status of chances.

In this section I consider more broadly whether there are any reasons to favor Demarest's interpretation, in particular in the sense of the just given second characterization of the issue. The question is whether the world should be thought to have an initial chance event, when one might consider that it is chancy in various other ways, e.g. its laws of evolution themselves are always probabilistically indeterministic.

First of all, it is worth mentioning that from the point of view given by the contemporary standard model of cosmology this question is moot. The so-called ΛCDM model, a development of the older standard big bang model, is a model of the general theory of relativity, a theory which makes use of no probabilities at all in its basic description of gravitating systems (including the universe). In this different sense it is also true that the universe never had a chance.

Demarest is not particularly interested in cosmology or the universes of general relativity however. She is concerned with probabilistic theories like classical statistical mechanics and quantum mechanics as applied to the world at large. We should, that is, imagine a statistical mechanical universe or a quantum mechanical universe (never minding that no concrete such model exists in physics that describes our universe) as a conceptual possibility when asking metaphysical questions about the world. Given the different ways of implementing probabilities in such a universe, we should ask whether one way is preferable to the others.

I should point out that this is not Demarest's question, for she explicitly restricts attention to "deterministically evolving worlds". Of course these worlds are not actually deterministic so long as the probabilities involved are chances. Nevertheless, unaffected by that fact is one of her central points: "that positing just one initial chance event can justify the usefulness and explain the ubiquity of nontrivial probabilities to epistemic agents like us, even if there are no longer any chance events in our world"(Demarest, 2016, 249). I say: so can a lot of other ways of conceiving chance in these theories. It is therefore necessary to compare them if we are to take Demarest's (and Loewer's) account seriously.

For present purposes, I am happy to agree with Demarest that the initial chance account can indeed justify and explain nontrivial probabilities used to describe subsystems of the universe.[4] But is it a good explanation? Is it worth believing?

[4]Notwithstanding pressure to move in this "global" direction in statistical mechanics (Callender, 2011)

The initial chance account invites the oft-invoked (in cosmology) picture of the (blind and unskilled) Creator throwing a dart (Wald, 2006, 396) or pointing a pin (Penrose, 1989, 442) at the set of possible universes, thereby picking out the initial conditions of the universe. That such pictures are intended as pejorative jabs at dubious metaphysics is plain. A mere picture is hardly an objection, of course, so what is it that seems problematic about initial chances for the universe? Could it not be the best cosmogonical story of our universe, that is, that a matter of chance determined its actualization out of a vast range of possibilities that could have been actualized had only their sisal been struck?

Intuition suggests that this just is not a serious, satisfying story for how the world could be. The probabilities of events in the actual world would derive ultimately from the probabilities for the actualization of our world. But why should we not just assume that the world started in the state that it did, with probability one or with certainty? Presumably the response of the initial chance advocate is that in that case we would lose the justification and explanation of subsystem probabilities. Yet is there anything to lose, if this metaphysical explanation is epistemically untrustworthy? How can we come to know these ultimate probabilities of other worlds? Is the metaphysical story sufficiently complete even? How could the probabilities of other worlds matter for what happens in *our* world?

I am willing to grant that these questions do have some answer, for what strikes me as a more serious difficulty is the following. Insofar as they are objective and justified, the probabilities agents like us use for specific events in subsystems of the world must be epistemic probabilities. On Demarest's (and Loewer's) account all such epistemic probabilities derive from initial epistemic probabilities for different initial conditions of the world. How is it that these probabilities obtain their needed objectivity and justification, and hence explanatory power? According to Demarest it is because they accord with the actual chances. However, what has one achieved by invoking "actual chances" at this stage? Although these chances do not merely have a *virtus dormitiva* per se, "just so" stories like this surely make the explanatory credentials of chances suspect. Does one dare invoke a transcendental argument or thump the realist table to defend their objectivity?

If we were somehow forced to adopt the initial chance explanation of epistemic probabilities, then we might swallow whatever dubious metaphysics attendant to it. If there were reasonable alternatives, however, should we not prefer them? And indeed there are other interpretive options available. Locating the chances at another time (or even "outside the universe") constitutes one set of possibilities, but they obviously suffer from the same awkwardness as the initial chance account. Another is based on the idea that chancy behavior occurs at discrete time intervals. One finds this idea in the orthodox Copenhagen and other collapse interpretations of quantum mechanics for example. One might be uneasy with the invocation of chancy behavior at potentially ill-defined times in such interpretations, and even with their postulation of two dynamical laws of nature, a deterministic one and an indeterministic one (although it is a feature of the initial chance account as well). However one at least avoids a commitment to chance figuring into

(and quantum mechanics) in order to justify and explain probabilities in subsystems of the universe, serious reservations about whether doing so is itself justified are advanced by, inter alia, Earman (2006).

cosmogenesis and also the questionable leap to objectivity in agential probabilities, since chances in these interpretations are physical processes that happen within the universe, whether as part of the general evolution of the universe or tied to the evolution of individual systems.

Another possibility is suggested by continuing this line of thought, i.e. of spreading chanciness out further in time. Instead of chancy behavior at discrete intervals, why not suppose that it occurs continuously? In quantum mechanics this idea is implemented in some interpretations, such as continuous spontaneous localization, and in statistical mechanics there are various stochastic dynamics approaches. Advantages of this idea are that one has a single law of evolution, an indeterministic one, and, again, one does not make chanciness a matter of cosmogenesis. What disadvantage? To some that it makes the world rife with indeterminism. Yet who is afraid of indeterminism? It surely does not mean anything goes, nor does it threaten the possibility of knowledge of the world (although there are limits to what we can know). Besides, by accepting quantum mechanics (or even statistical mechanics) we have already let indeterminism in the door in physics.

When we look at the interpretations available for a world governed by probabilistic laws, in every case the alternatives to the initial chances view therefore appear preferable. Indeed, it would seem that only one who demands that the world be as deterministic as possible could favor the initial chances view, but it is hard to see what motivation there could be for that demand. I therefore conclude, in a final sense, that *the universe never had a chance*.

That said, I emphasize that this judgment applies only to the case where we treat the universe as a statistical mechanical system or quantum mechanical system. In other words, the world is the universe, our world-metaphysics is our universe-metaphysics. The considerations leading to this conclusion change shape somewhat when we confine the application of our theories to systems describable by those theories. The initial chance account is far less dubious when attached to individual statistical mechanical systems and not automatically to the universe at large. Indeed, it could well be that the initial conditions of similar systems are best treated as randomly distributed, for here we do have empirical evidence that this interpretation can be used to explain—unlike with the universe, where we have but one system.

There is, as noted, sometimes pressure to globalize our theories, especially in the case of statistical mechanics. If we ask what accounts for the randomness in initial conditions of a particular class of systems, it is natural to look at larger systems that contain them. If we find that these systems have random initial conditions, then we continue to expand our scope, ultimately reaching the "maximally inclusive entity whose parts are all the things that exist." This globalization of statistical mechanics is the kernel of the so-called imperialism of (Albert, 2000) and Loewer. If we are right to feel this pressure to interpret the world at large in the same terms as individual physical systems, then there is concomitant pressure to hold the same interpretive of chance in both cases. I have argued, however, that the intuitive considerations vary somewhat, at least with respect to the initial chance account. Is this reason to disfavor it in the case of individual systems? Or is our confidence in its applicability for individual systems sufficient to overcome any hesitation at

accepting it for the universe? My inclination is to answer "yes" and "no", but I offer no grounds for the preference here. I do believe that metaphysicians of science should care about considerations like this, however, having to do with the relation of subsystem and universe, for often enough what seems right in one context is questionable in the other.

I close this section with a brief comment on the relation of Loewer's and Demarest's accounts. As I argued above, empirical evidence and explanatory considerations do not favor one over the other, since they account for empirical evidence in essentially the same way. The central difference is whether chances are understood as reducible to other facts, hence not part of the fundamental ontology of the world, or as "robustly metaphysical", in which case they are. The problems Demarest mentions for the Humean view—past events may have nontrivial chances, the chance of an event depends on what one knows, worlds with identical frequencies cannot have different chances, etc.—are surely not problems when viewed properly through the Humean lens. However, whereas the problem I raise for the initial chance view, concerning the explanatory credentials and justification for the posit of initial chances, threatens Demarest's account, it will not worry the Humean of Loewer's stripe, for these initial chances do not exist for the Humean. Humean chances do not produce or generate any actual states of affairs. Of course one may raise the usual complaint against the Humean, that there is a circularity in the Humean account involving descriptions explaining themselves, and others besides. I do not care to enter into this debate here of course. I only wish to point out that my argument about how chance can fit into a cosmogonical worldview appears to give some reason to favor the Humean account in this particular context.

4 Conclusion

In this paper I considered whether we should think that the world had one chance, as claimed by Demarest. First I considered her claim that we have good evidence that an initial chance event occurred by contrasting it with relevant classes of alternatives. I argued that evidence neither favors a chancy theory over a chanceless theory, nor initial chances over other implementations of chances, nor metaphysically robust chances over Humean chances. I concluded, therefore, that we do not have good evidence to adopt the initial chance account.

I then considered whether there were other reasons to favor or disfavor the initial chance account. I argued that the dubious nature of worldly chances provides a strong impulse to look for other accounts that do not make chance a matter of cosmogenesis. The other implementations did not suffer from this defect, so I suggested that from a cosmogonical perspective they should be preferred. But the relation of the universe and its subsystems makes a demand to have a consistent interpretation. As the initial chance account looks favorable on the subsystem level (to many) and not on the universe's level (as I argued), there remains a significant metaphysical tension to be resolved.

References

Albert, D. (2000). *Time and Chance*. Cambridge, MA: Cambridge, MA: Harvard University Press.

Callender, C. (2011). The past histories of molecules. In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 83–113. Oxford: Oxford University Press.

Demarest, H. (2016). The universe had one chance. *Philosophy of Science 83*(2), 248–264.

Earman, J. (1986). *A Primer on Determinism*. Dordrecht: D. Reidel Publishing Company.

Earman, J. (1995). *Bangs, Crunches, Whimpers, and Shrieks*. Oxford: Oxford University Press.

Earman, J. (2006). The "past hypothesis": Not even false. *Studies in History and Philosophy of Modern Physics 37*, 399–430.

Earman, J. and J. Norton (1987). What price spacetime substantivalism? the hole story. *British Journal for the Philosophy of Science 38*, 515–525.

Harrison, E. (2000). *Cosmology: the science of the universe* (2nd ed.). Cambridge: Cambridge University Press.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Lewis, D. (1986). *Philosophical Papers*, Volume 2. Oxford: Oxford University Press.

Loewer, B. (2001). Determinism and chance. *Studies in History and Philosophy of Modern Physics 32*, 609–620.

Lyon, A. (2011). Deterministic probability: neither chance nor credence. *Synthese 182*, 413–432.

Norton, J. (2008). The dome: An unexpectedly simple failure of determinism. *Philosophy of Science 75*, 786–798.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

Schaffer, J. (2010). Monism: The priority of the whole. *The Philosophical Review 119*, 31–76.

Wald, R. (2006). The arrow of time and the initial conditions of the universe. *Studies in History and Philosophy of Modern Physics 37*, 394–398.

Werndl, C. (2009). Are deterministic descriptions and indeterministic descriptions observationally equivalent? *Studies in History and Philosophy of Modern Physics 40*, 232–242.

Winsberg, E. (2008). Laws and chances in statistical mechanics. *Studies in History and Philosophy of Modern Physics 39*, 872–888.

Wüthrich, C. (2011). Can the world be shown to be indeterministic after all? In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 365–389. Oxford: Oxford University Press.

Draft paper for the symposium *Mechanism Meets Big Data: Different Strategies for Machine Learning in Cancer Research* to be held at *PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association* (Seattle, WA; 1-4 Nov 2018).

MECHANISTIC MODELS AND THE EXPLANATORY LIMITS OF MACHINE LEARNING

Emanuele Ratti[1], University of Notre Dame

Ezequiel López-Rubio, Universidad Nacional de Educación a Distancia, University of Málaga

Abstract

We argue that mechanistic models elaborated by machine learning cannot be explanatory by discussing the relation between mechanistic models, explanation and the notion of intelligibility of models. We show that the ability of biologists to understand the model that they work with (i.e. intelligibility) severely constrains their capacity of turning the model into an explanatory model. The more a mechanistic model is complex (i.e. it includes an increasing number of components), the less explanatory it will be. Since machine learning increases its performances when more components are added, then it generates models which are not intelligible, and hence not explanatory.

1. INTRODUCTION

Due to its data-intensive turn, molecular biology is increasingly making use of machine learning (ML) methodologies. ML is the study of generalizable extraction of patterns from data sets starting from a problem. A problem here is defined as a given set of input variables, a set of outputs which have to be calculated, and a sample (previously input-output pairs already observed). ML calculates a quantitative relation between inputs and outputs in terms of a predictive model by learning from an already structured set of input-output pairs. ML is expected to increase its performances when the complexity of data sets increase, where complexity refers to the number of input variables and the number of samples. Due to this capacity to handle complexity, practitioners think that ML is potentially able to deal with biological systems at the macromolecular level, which are notoriously complex. The development of ML has been proven useful not just for the

[1] mnl.ratti@gmail.com

complexity of biological systems *per se*, but also because biologists now are able to generate an astonishingly amount of data. However, we claim that the ability of ML to deal with complex systems and big data comes at a price; *the more ML can model complex data sets, the less biologists will be able to explain phenomena in a mechanistic sense*.

The structure of the paper is as follows. In Section 2, we discuss mechanistic models in biology, and we emphasize a surprising connection between explanation and model complexity. By adapting de Regt's notion of pragmatic understanding (2017) in the present context, we claim that if a how-possibly mechanistic model can become explanatory, then it must be intelligible to the modeler (Section 2.2, 2.3 and 2.4). Intelligibility is the ability to perform precise and successful material manipulations on the basis of the information provided by the model about its components. The results of these manipulations are fundamental to recompose the causal structure of a mechanism out of a list of causally relevant entities. Like a recipe, the model must provide instructions to 'build' the phenomenon, and causal organization is fundamental in this respect. If a model is opaque to these organizational aspects, then no mechanistic explanations can be elaborated. By drawing on studies in cognitive psychology, we show that the more the number of components in a model increases (the more the model is complex), the less the model is intelligible, and hence the less an explanation can be elaborated.

Next, we briefly introduce ML (Section 3). As an example of ML application to biology, we analyze an algorithm called PARADIGM (Vaske et al 2010), which is used in biomedicine to predict clinical outcomes from molecular data (Section 3.1). This algorithm predicts the activities of genetic pathways from multiple genome-scale measurements on a single patient by integrating information on pathways from different databases. By discussing the technical aspects of this algorithm, we will show how the algorithm generates models which are more accurate as the number of variables included in the model increases. By variables, here we mean biological entities included in the model and the interactions between them, since those entities are modeled by variables in PARADIGM.

In Section 4 we will put together the results of Section 2 and 3. While performing complex localizations more accurately, we argue that an algorithm like PARADIGM makes mechanistic models so complex (in terms of the number of model components) that no explanation can be constructed. In other words, ML applied to molecular biology undermines biologists' explanatory abilities.

2. COMPLEXITY AND EXPLANATIONS IN BIOLOGY

The use of machine learning has important consequences for the explanatory dimension of molecular biology. Algorithms like PARADIGM, while providing increasingly accurate localizations, challenge the explanatory abilities of molecular biologists, especially if we assume the account of explanation of the so-called mechanistic philosophy (Craver and Darden 2013; Craver 2007; Glennan 2017). In order to see how, we need to introduce the notion of mechanistic explanation, and its connection with the notion of intelligibility (de Regt 2017).

2.1 Mechanistic explanations

Molecular biology's aim is to explain how phenomena are produced and/or maintained by the organization instantiated by macromolecules. Such explanations take the form of mechanistic descriptions of these dynamics. As Glennan (2017) succinctly emphasizes, mechanistic models (often in the form of diagrams complemented by linguistic descriptions) are vehicles for mechanistic explanations. Such explanations show how a phenomenon is produced/maintained and constituted by a mechanism – mechanistic models explain by explaining *how*. As Glennan and others have noticed, a mechanistic description of a phenomenon looks like what in historical narrative is called *causal narrative*, in the sense that it "describes sequences of events (which will typically be entities acting and interacting), and shows how their arrangement in space and time brought about some outcome" (Glennan 2017, p 83). The main idea is that we take a set of entities and activities to be causally relevant to a phenomenon, and we explain the phenomenon by showing how a sequence of events involving the interactions of the selected entities produces and/or maintains the explanandum. In epistemic terms, it is a

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -40-

matter of showing a chain of inferences that holds between the components of a model (e.g. biological entities). Consider for instance the phenomenon of restriction in certain bacteria and archaea (Figure 1). This phenomenon has been explained in terms of certain entities (e.g. restriction and modification enzymes) and activities (e.g. methylation). Anytime a bacteriophage invades one of these bacteria or archaea (from now on *host cells*), host cells stimulate the production of two types of enzymes, i.e. a restriction enzyme and a modification enzyme. The restriction enzyme is designed to recognize and cut specific DNA sequences. Such sequences, for reasons we will not expose here[2], are to be found in the invading phages and/or viruses. Hence, the restriction enzyme destroys the invading entities by cutting their DNA. However, the restriction enzyme is not able to distinguish between the invading DNA and the DNA of the host cell. Here the modification enzyme helps, by methylating the DNA of the host cell at specific sequences (the same that the restriction enzyme cuts), thereby preventing the restriction enzyme to destroy the DNA of the host cell. The explanation of the phenomenon of restriction is in terms of a narrative explaining how certain entities and processes contribute to the production of the phenomenon under investigation. The inferences take place by thinking about the characteristics of the entities involved, and how the whole functioning of the system can be recomposed from entities themselves.

[2] See for instance (Ratti 2018)

Seattle, WA; 1-4 November 2018 -41-

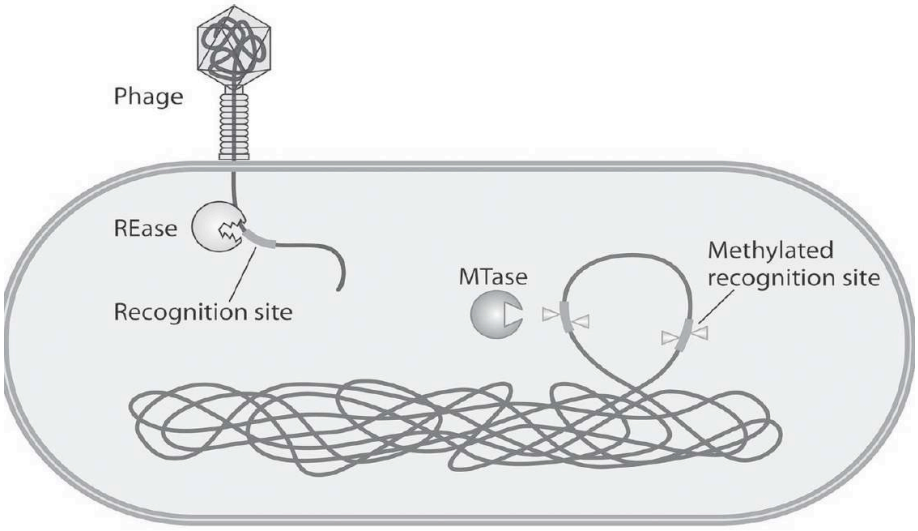


Figure 1. Mechanistic model of restriction. A phage enters a bacterium cell and sequences of its DNA are cleaved by a restriction enzyme (REase). Simultaneously, a modification enzyme (MTase) methylates a specific sequence in the DNA of host so that the restriction enzyme does not cleave the genome of the host too. Original figure taken from (Vasu and Nagaraja 2013).

2.2. Complexity of mechanistic models

Despite the voluminous literature on mechanistic explanation, there is a connection between models, *in fieri* explanations and the modeler that has not been properly characterized. In particular, mechanistic models should be intelligible to modelers in order to be turned into complete explanations. Craver noticed something like that when he states that his ideal of completeness of a mechanistic description (in terms of molecular details) should not be taken literary, but completeness always refer to the particular explanatory context one is considering. The reason why literary completeness is unattainable is because complete models will be of *no use* and completely *obscure* to modelers; "such descriptions would include so many potential factors that they would be *unwieldy for the purpose of prediction and control and utterly unilluminating to human beings*" (2006, p 360, emphasis added).

We rephrase Craver's intuitions by saying that *how-possibly models cannot be turned into adequate explanations if they are too complex*. We define complexity as *a function of the number of entities and activities (i.e. components of the model) that have*

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -42-

to be coordinated in an organizational structure in the sense specified by mechanistic philosophers. This means that no agent can organize the entities and/or activities localized by highly complex models in a narration that rightly depicts the organizational structure of the *explanandum*. Therefore, very complex models which are very good in localization cannot be easily turned into explanations. Let us show why complex models cannot be turned into explanatory models in the mechanistic context.

2.3 Intelligibility of mechanistic models

The idea that agents cannot turn highly complex mechanistic models into explanations can be made more precise by appealing to the notion of *intelligibility* (de Regt 2017).

By following the framework of models as mediators (Morgan and Morrison 1999), de Regt argues that models are the way theories are applied to reality. Similar to Giere (2010), de Regt thinks that theories provide principles which are then articulated in the form of models to explain phenomena; "[t]he function of a model is to represent the target system in such a way that the theory can be applied to it" (2017, p 34). He assumes a broad meaning of explanation, in the sense that explanations are arguments, namely attempts to "answer the question of why a particular phenomenon occurs or a situation obtains (…) by presenting a systematic line of reasoning that connects it with other accepted items of knowledge" (2017, p 25). *Ça va sans dire*, arguments of the sort are not limited to linguistic items[3]. On this basis, de Regt's main thesis is that a *condition sine qua non* to elaborate an explanation is that the theory from which it is derived must be intelligible.

In de Regt's view, the intelligibility of a theory (*for scientists*) is "[t]he value that scientists attribute to the cluster of virtues (…) that facilitate the use of the theory for the construction of models" (p 593). This is because an important aspect of obtaining explanations is to derive models from theories, and to do that a scientist must use the theories. Therefore, if a theory possesses certain characteristics that make it easier to be used by a scientist, then the same scientist will be in principle more successful in deriving explanatory models. In (2015) de Regt extends this idea also to models in the sense that "understanding consists in being able to use and manipulate the model in order to make

[3] Mechanistic explanations are arguments, though not of a logical type

inferences about the system, to predict and control its behavior" (2015, p 3791). If for some reasons models and theories are not intelligible (to us), then we will not be able to develop an explanation, because we would not know how to use models or theories to elaborate one.

This idea of intelligibility of models and its tight connection with scientific explanation, can be straightforwardly extended to mechanistic models. Intelligibility of mechanistic models is defined by the way we *successfully* use them to explain phenomena. But how do we use models (mechanistic models in particular), and for what? Please keep in mind that whatever we do with mechanistic models, it is with explanatory aims in mind. Anything from predicting, manipulating, abstracting, etc is because we want an explanation. This is a view shared both by mechanistic philosophers but by de Regt as well, whose analysis of intelligibility is in explanatory terms.

First, highly abstract models can be used to build more specific models, as in the case of schema (Machamer et al 2000; Levy 2014). A schema is "a truncated abstract description of a mechanism that can be filled with descriptions of known component parts and activities" (Machamer et al 2000, p 16). For instance, consider the model of transcription. This model can be highly abstract where 'gene' stands for any gene, and 'transcription factor' stands for any transcription factor. However, we can instantiate such a schema in a particular experimental context by specifying which gene and which transcription factors are involved. The idea is that biologists, depending on the specific context they are operating, can instantiate experiments to find out which particular gene or transcription factor is involved in producing a phenomenon at a given time.

Next, mechanistic models can be used in the context of the *build-it test* (Craver and Darden 2013) with confirmatory goals in mind. Since mechanistic explanations may be understood as recipes for construction, and since recipes provide instructions to use a set of ingredients and instruments to produce something (e.g. a cake), then mechanistic models provide instructions to build a phenomenon or instructions to modify it in controlled ways because, after all, they tell us about the internal division of labor between entities causally relevant to producing or maintaining phenomena. This is in essence the build-it test as a confirmation tool; by modifying an experimental system on the basis of the 'instructions' provided by the model that allegedly explains such a phenomenon, we

get hints as to how the model is explanatory. If the hypothesized modifications produce in the 'real-world' the consequences we have predicted on the basis of the model, then the explanatory adequacy of the model is corroborated. The more the modifications suggested are precise, the more explanatory the model will be[4]. A first lesson we can draw is that *if a mechanistic model is explanatory, then it is also intelligible*, because it is included in the features of being explanatory mechanistically the fact that we can use the model to perform a build-it test.

The build-it test is also useful as a *tool to develop* explanations. Consider again the case of restriction in bacteria and a how-possibly model of this phenomenon based on a few observations. Let's say that we have noticed that when phages or viruses are unable to grow in specific bacteria, such bacteria also produce two types of enzymes. We know that the enzymes, the invading phages/viruses and restriction are correlated. The basic model will be as follows; anytime a phage or a virus invade a bacterium, these enzymes are produced, and hence the immune system of the bacterium must be related to these enzymes. We start then to instantiate experiments on the basis of this simple model. Such a model suggests that these enzymes must do something to the invading entities, but that somehow modify the host cell as well. Therefore, the build-it test would consist in a set of experiments to stimulate and/or inhibit these entities to develop our ideas about the nature of their causal relevance and their internal division of labor. *In fieri* mechanistic models suggest a range of instructions to 'build' or 'maintain' phenomena. These instructions are used to instantiate experiments to refine the model and make it explanatory. This is an example of what Bechtel and Richardson would call *complex localization* (2010, Chapter 6), and it is complex because the strategy used to explain the behavior of a system (immune system of host cells) is heavily constrained by empirical results of lower-levels. The how-possibly model affords a series of actions leading to a case of complex localization, when "constraints are imposed, whether empirical or theoretical, they can serve simultaneously to vindicate the initial localization and to develop it into a full-blooded mechanistic explanation" (Bechtel and Richardson 2010, p 125). Therefore, *if a how-possibly model can be turned into an explanatory model, then it*

[4] Please note that such a test, when involving adequate mechanistic explanations, is also the preferred way to teach students in text books, or also a way to provide instructions to reproduce the results of a peer-reviewed article

Seattle, WA; 1-4 November 2018 -45-

is intelligible, because the way we turn it into an explanatory model is by instantiating build-it tests.

A mechanistic model is therefore intelligible either when (a) it is a schema and we can instantiate such a model in specific contexts, or (b) when it affords a series of built-it test which are used either to corroborate its explanatory adequacy, or to make it explanatory. About (b), it should be noted that if we consider a mechanistic model as a narrative, then the model will be composed of a series of steps which influence each other in various ways. *Being able to use a model means being able to anticipate what would happen to other steps if I modify one step in particular*. This is not a yes/no thing. The model of restriction-modification systems is highly intelligible, because I know that if I prevent the production of modification enzymes I simultaneously realize that the restriction enzyme will destroy the DNA of the host cell. However, more detailed models will be less intelligible, because it would be difficult to simultaneously anticipate what would happen at each step by modifying a step in particular.

2.4 Recomposing mechanisms and intelligibility

In the mechanistic literature, the process of developing an explanatory model out of a catalogue of entities that are likely to be causally relevant to a phenomenon is called *recomposition of a mechanism* and it usually happens after a series of localization steps.

To recompose a mechanism, a modeler must be able to identify causally relevant entities and their internal division of labor. The idea is not just to 'divide up' a given phenomenon in tasks, but also a given task in subtasks interacting in the overall phenomenon, as it happens in complex localization (Bechtel and Richardson 2010). In the simplest case, researchers assume linear interactions between tasks, but there may be also non-linear or more complex type of interactions.

These reasoning strategies are usually implemented by thinking about these dynamics with the aid of *diagrams*. Diagrammatic representations usually involve boxes standing for entities (such as genes, proteins, etc) and arrows standing for processes of various sorts (phosphorylation, methylation, binding, releasing, etc). Therefore, biologists recompose mechanisms as mechanistic explanations by thinking about these diagrams,

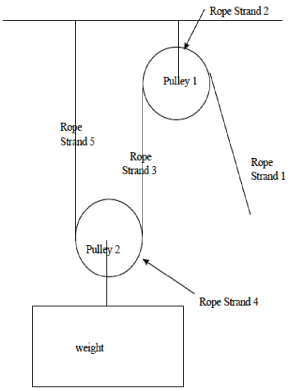
and they instantiate experiments (i.e. built-it test) exactly on the basis of such diagrammatic reasoning.

Cognitive psychology and studies of scientific cognition have extensively investigated the processes of diagrammatic reasoning (Hegarty 2000; 2004; Nersessian 2008). Moreover, empirical studies have emphasized the role of diagrams in learning and reasoning in molecular biology (Kindfield 1998; Trujillo 2015). In these studies, diagrammatic reasoning is understood as a "task that involves inferring the behavior of a mechanical system from a visual-spatial representation" (Hegarty 2000, p 194). Hegarty refers to this process as *mental animation*, while Nersessian (2008) thinks about this as an instantiation of *mental modelling*. This is analogous to thinking about mechanistic models as narratives, namely being able to infer how a course of events, decomposed into steps, may change if we change one step in particular. Mental animation is a process of complex visual-spatial inference. Limits and capabilities of humans in such tasks depend on the cognitive architecture of human mind[5]. What Hegarty has found is that mental animation is *piecemeal*, in the sense that human mind does not animate the components of a diagram in parallel, but rather infer the motion of components *one by one*. This strategy has a straightforward consequence; in order to proceed with animating components, we should store intermediate results of inferences drawn on previous components. Due to the limitations of working memory (WM), people usually store such information on external displays. Hegarty has provided evidence that diagrammatic reasoning is bounded to WM abilities. The more we proceed in inferring animation on later components, the more the inferences on earlier components degrade (see for instance Figure 2); "as more components of the system are 'read into' spatial working memory, the activation of all items is degraded, so that when later components are in, there is not enough activation of the later components to infer their motion" (Hegarty 2000, p 201).

[5] On this, I rely on the framework assumed by the cognitive-load theory (Paas et al 2010)

Seattle, WA; 1-4 November 2018 -47-

a



b

Cycle	Action Taken
1	Rope Strand 1 is inspected (simulation is given knowledge that this rope strand is being pulled)
2	Rope Strand 1 is animated (i.e. its direction of motion is inferred)
3	Rope Strand 2 is inspected
4	Rope Strand 2 is animated
5	Rope Strand 3 is inspected Pulley 1 is inspecte
6	Rope Strand 3 is animated Pulley 1 is animated
7	Rope Strand 4 is inspected
8	Rope Strand 4 is animated
9	Rope Strand 5 is inspected Pulley 2 is inspected
10	Rotation of Pulley 2 is inferred Translation of Pulley 2 is inferred
11	Weight is inspected
12	Weight is animated

Figure 2. (a) Example of diagram of a simple pulley system that can be mentally animated (b) Description of typical actions that can be one by one to animate the pulley system. Both figures taken from Hergaty (2000)

The actual limit of our cognitive architecture on this respect may be debated, and it is an empirical issue. The important point is that *no matter our external displays*, for very large systems (such as Figure 3) it is very unlikely that human cognition will be able to process all information about elements interactivity. This is because by animating components one-by-one, even if we use sophisticated instruments such computer simulations, still inferences on earlier components will degrade. This means that build-it tests will be very ineffective, if not impossible. In terms of narratives, recipes and mechanistic models, this means that for large mechanistic diagrams with many model components, no human would be able to anticipate the consequences of modifying a step in the model for all the other steps of the model, even if a computer simulation shows that the phenomenon can be possibly produced by the complex model. The computer simulation may highlight certain aspects (as Bechtel in 2016 notes), but the model is not intelligible in the sense required by mechanistic philosophy. *If the model is not intelligible in this way, then it cannot be possibly turned into an explanation*.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -48-

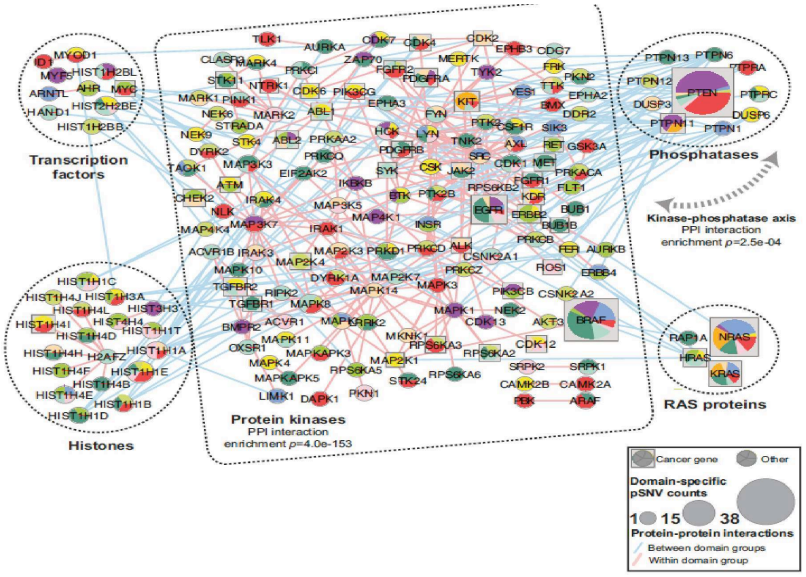


Figure 3. Network of interactions of proteins with significant enrichment of phosphorylation-related single nucleotide variations. Phosphorylation is a central post-translational modification in cancer biology. Authors are not trying to recompose the mechanism that from phosphorylated proteins (nodes) lead to a tumor phenotype, but rather to identify the magnitude of the impact of this process on cancer genes. Figure taken from (Reimand et al 2013)

The results of Hegarty's research suggest that when mechanistic models are concerned, strategies of localization are effective (in terms of explanatory potential) only when a limited number of model components are actually identified. The number may increase if we use computer simulations. However, for very large amounts of model components (such as Figure 3) recomposition is just impossible for humans, because inferences on the role of components in the causal division of labor of a phenomenon will degrade to make place for inferences about other components. This of course holds only if we have explanatory aims in mind.

To summarize, in section 2 we have made three claims:

1. If a how-possibly model can be turned into an explanation, then it is intelligible
2. If a model is not intelligible, then it cannot become explanatory
3. Complex models are a class of non-intelligible models

3. MACHINE LEARNING AND LOCALIZATIONS

Machine learning (ML) is a subfield of computer science which studies the design of computing machinery that improves its performance as it learns from its environment. A ML algorithm extracts knowledge from the input data, so that it can give better solutions to the problem that it is meant to solve. This learning process usually involves the automatic construction and refinement of a model of the incoming data. In ML terminology, a model is an information structure which is stored in the computer memory and manipulated by the algorithm.

As mentioned before, the concept of 'problem' in ML has a specific meaning which is different from other fields of science. A ML problem is defined by a set of input variables, a set of output variables, and a collection of samples which are input-output pairs. Solving a problem here means finding a quantitative relation between inputs and outputs in the form of a predictive model, in the sense that the algorithm will be used to produce a certain output given the presence of a specific input.

3.1 The PARADIGM algorithm

ML has been applied in the molecular sciences in many ways (Libbrecht and Noble 2015). Especially in cancer research[6], computer scientists have created and trained a great deal of algorithms in order to identify entities that are likely to be involved in the development of tumors, how they interact, to predict phenotypes, to recognize crucial sequences, etc (see for instance Leung et al 2016).

As a topical example of ML applied to biology, we introduce an algorithm called PARADIGM (Vaske et al 2010). This algorithm is used to infer how genetic changes in a patient influence or disrupt important genetic pathways underlying cancer progression. This is important because there is empirical evidence that "when patients harbor genomic alterations or aberrant expression in different genes, these genes often participate in a common pathway" (Vaske et al 2010, p i237). Because pathways are so large and biologists cannot hold in their mind the entities participating in them, PARADIGM integrate several genomic datasets – including datasets about interactions between genes and phenotypic consequences – to infer molecular pathways altered in patients; it predicts

[6] See for instance The Cancer Genome Atlas at https://cancergenome.nih.gov

whether a patient will have specific pathways disrupted given his/her genetic mutations.

The algorithm is based on a simplified model of the cell. Each biological pathway is modeled by a graph. Each graph contains a set of nodes, such that each node represents a cell entity, like a mRNA, a gene or a complex. A node can be only in three states (i.e. activated, normal or deactivated). The connections among nodes are called factors, and they represent the influence of some entities on other entities. It must be noticed that the model does not represent why or how these influences are exerted. Only the sign of the influence, i.e. positive or negative, is specified.

The model specifies how the expected state of an entity must be estimated. The entities which are connected by positive or negative factors to the entity at hand cast votes which are computed by multiplying +1 or –1 by the states of those entities, respectively. In addition to this, there are 'maximum' and 'minimum' connections to cast votes which are the maximum or the minimum of the states of the connected entities, respectively. Overall, the expected state of an entity is computed as the result of combining several votes obtained from the entities which are connected to it. Such a voting procedure can be associated to localizations (i.e. whether a node is activated or not), but hardly to biological explanations.

The states of the entities can be hidden, i.e. they can not be directly measured on the patients, or observable. The states of the hidden variables must be estimated by a probabilistic inference algorithm, which takes into account the states of the observed variables and the factors to estimate the most likely values of the hidden variables. Here it must be pointed out that this algorithm does not yield any explanation about the computed estimation. Moreover, it could be the case that the estimated values are not the most likely ones, since the algorithm does not guarantee that it finds the globally optimum solution.

The size of the model is determined by the number of entities and factors that the scientist wishes to insert. A larger model provides a perspective of the cell processes which contains more elements, and it might yield better predictions. This means that the more components the model has, the better the algorithm will perform. In biological terms, the larger the model, the more precise *complex localizations* the algorithm will identify, in particular by pointing more precisely towards pathways that are likely to be

disrupted in the patient with more information about the state of gene activities, complexes and cellular processes. Importantly, PARADIGM does not infer new genetic interactions, but it just helps identifying those known interaction in a new data set. It is completely supervised, in the sense that "[w]hile it infers hidden quantities (…), it makes no attempt to infer new interactions not already present in an NCI [National Cancer institute database] pathway" (Vaske et al 2010, p i244).

4 COMPLEX MODELS AND MECHANISTIC EXPLANATIONS

Before unwinding our conclusions, let me recall the results of Section 2 very briefly:

1. If a how-possibly model can be turned into an explanation, then it is intelligible[7].
2. If a model is not intelligible, then it cannot become explanatory
3. Complex model (in the sense explained in 2.2) are not intelligible

What does this have to do with PARADIGM? It is important to emphasize what we have pointed out in Section 3.1, namely that an algorithm like PARADIGM is more efficient when working with more components. If we think about models generated by algorithms such as PARADIGM in mechanistic terms, this means that the algorithm provides more precise complex localizations, because more entities that are likely to be causally relevant to a phenomenon are identified, and the information about the probability of a pathway being disrupted in a patient will be more precise. However, the models will be more complex, and they will be decreasingly intelligible. This is because the final model will count an elevated number of components, and recomposing these components into a full-fledged mechanistic explanation of how a tumor is behaving will be cognitively very difficult; the inferences about the behavior of components are not run in parallel, but one by one, and once we proceed in inferring the behavior of a component on the basis of the behavior of another component, other inferences will degrade, as Hegarty's studies have shown. In the ideal situation, PARADIGM will generate unintelligible models:

[7] Remember: A mechanistic model *x* is intelligible to a modeler *y* if *y* can use the information about the components of *x* to instantiate so-called 'build-it test'. Such tests are performed on how-possibly models to turn them into explanatory models by obtaining information on how to recompose a phenomenon (i.e. by showing how a list of biological entities are organized to produce a phenomenon).

4. Algorithms such as PARADIGM generate models which are not intelligible because such models are too complex
5. Because of 2, 3 and 4, complex models generated out of algorithms like PARADIGM cannot become explanations

This means that when we use algorithms such as PARADIGM to cope with the complexity of biological systems, we successfully handle big data sets, but such a mastery comes at a price. Using ML in molecular biology means providing more detailed localizations, but we also lose explanatory power, because no modeler will be able to recompose the mechanism out of a long list of entities.

This implies that, in the mechanistic epistemic horizon, the central role assigned to explanations should be reconsidered when contemporary molecular biosciences are concerned. As Bechtel has also emphasized in the context of computational models in mechanistic research (2016), such tools are useful to show whether some entities are likely to be involved in a particular phenomenon or suggest alternative hypotheses about the relation between certain entities. However, providing fully-fledged mechanistic explanations is another thing. It is the same with algorithms of ML; we identify more entities likely to be involved in a mechanism, we may even find out that entities involved in specific process may be connected with entities involved in other processes (via for instance Gene Ontology enrichments), but we cannot recompose a mechanism out of a list of hundreds of entities. In fact, we come to value different epistemic values, and *explanatory power is not one of them*. This somehow implies also a shift in the way scientific articles are organized; if in 'traditional' molecular biology evidence converges towards the characterization of a single mechanism, in data-intensive biology we make a list of entities that can be involved in a phenomenon, but we do not necessarily connect those entities mechanistically (Alberts 2012). Another strategy (Krogan et al 2015) – though motivated more by biologically rather than cognitive reasons – is to abstract from macromolecular entities and consider only aggregates of them in the form of networks; whether establishing network topology is providing a mechanistic explanation remains an open question.

REFERENCES

Alberts, B. (2012). The End of "Small Science"? *Science*, *337*(September), 1230529.

Bechtel, W. (2016). Using computational models to discover and understand mechanisms. *Studies in History and Philosophy of Science Part A*, *56*, 113–121.

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity - Decomposition and Localization as Strategies in Scientific Research*. Cambridge, Massachusetts, and London, England: The MIT Press.

Craver, C. (2007). *Explaining the Brain - Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. Chicago: The University of Chicago Press.

De Regt, H. (2017). *Understanding Scientific Understanding*. Oxford: Oxford University Press.

de Regt, H. W. (2015). Scientific understanding: truth or dare? *Synthese*, *192*(12), 3781–3797. http://doi.org/10.1007/s11229-014-0538-7

Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese*, *172*(2), 269–281.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford University Press.

Hegarty, M. (2000). Capacity Limits in Mechanical Reasoning. In M. Anderson, P. Cheng, & V. Haarslev (Eds.), *Diagrams 2000* (pp. 194–206). Springer-Verlag.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A., & Ideker, T. (2015). The Cancer Cell Map Initiative: Defining the Hallmark Networks of Cancer. *Molecular Cell*, *58*(4), 690–698.

Levy, A. (2014). What was Hodgkin and Huxley's achievement? *British Journal for the Philosophy of Science*, *65*(3), 469–492.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about Mechanisms. *Philosophy of Science*, (67), 1–25.

Morrison, M., & Morgan, M. (1999). Models as mediating instruments. In M. Morrison & M. Morgan (Eds.), *Models as Mediators*. Cambridge University Press.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: The MIT Press.

Ratti, E. (2018). "Models of" and "models for": On the relation between mechanistic models and experimental strategies in molecular biology. *British Journal for the Philosophy of Science*.

Reimand, J., Wagih, O., & Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Scientific Reports*, *3*.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., … Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, *26*(12), 237–245.

Vasu, K., & Nagaraja, V. (2013). Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiology and Molecular Biology Reviews*, *77*(1), 53–72.

The Roles of Possibility and Mechanism in Narrative Explanation

Abstract

There is a fairly longstanding distinction between what are called the *ideographic* as opposed to *nomothetic* sciences. The nomothetic sciences, such as physics, offer explanations in terms of the laws and regular operations of nature. The ideographic sciences, such as natural history (or, more controversially, evolutionary biology), cast explanations in terms of narratives. This paper offers an account of what is involved in offering an explanatory narrative in the historical (ideographic) sciences. I argue that narrative explanations involve two chief components: a possibility space and an explanatory causal mechanism. The presence of a possibility space is a consequence of the fact that the presently available evidence underdetermines the true historical sequence from an epistemic perspective. But the addition of an explanatory causal mechanism gives us a reason to favor one causal history over another; that is, causal mechanisms enhance our epistemic position in the face of widespread underdetermination. This is in contrast to some recent work that has argued against the use of mechanisms in some narrative contexts. Indeed, I argue that an adequate causal mechanism is always involved in narrative explanation, or else we do not have an explanation at all.

1. Introduction

The historical sciences (geology, paleontology, evolutionary biology, etc.)[1] are usually thought to deploy different explanatory strategies than the non-historical sciences (Turner 2007; Turner 2013). Whereas physics, say, seeks explanations given in terms of general laws and the like, the historical sciences seek to explain in terms of narratives. In this paper I will argue for a version of narrative explanation involving two chief components: possibility spaces and causal mechanisms. It has recently been argued that complex historical narratives (to be defined later) can't support explanations involving causal mechanisms (Currie 2014). I argue that this is mistaken. I'll go over some recent work on the history of abiogenesis research to support this contention.

The argument presented in this paper will defend two primary claims: (1) the conceptual structure of narrative explanations nearly always involves a space of alternative possibilities. This can be for either epistemic or ontological reasons. From and epistemic perspective possibility spaces are necessary on account of our position relative to the available evidence. That is, the available evidence radically underdetermines any particular causal history, and on the basis of that fact many possible histories appear compatible with what we know (see Gordon and Olson 1994, p. 15). Construed ontologically, a set of historical facts might involve a high degree of objective contingency—it might be the case that things really could have gone a number of different ways. For the purposes of this paper I remain silent with respect to this ontological aspect and defend the importance of possibility spaces for largely epistemic reasons. (2) Adequate causal mechanisms enhance our epistemic position relative to alternative causal

[1] I note that the idea of evolutionary biology as a properly "historical science" is a controversial one. See Ereshefsky (1992) for some strong arguments against the idea of evolutionary biology as having a distinctively 'historical' flavor.

histories. Causal mechanisms put us in a position to better assess the plausibility of a given history within our possibility space, and in this way enhance the epistemic power of a purportedly explanatory historical narrative. This can involve either the actual discovery of such mechanisms, or raw theoretical innovation. Citing an adequate causal mechanism may not discriminate between possibilities in decisive fashion. Rampant underdetermination seems to rule out such a possibility (see Turner 2007). But an adequate mechanism does make a given explanation more explanatory than its competitors, and so part of the task is to see how this notion of mechanistic adequacy can be cashed out in such a way as to make this notion of *explanatoriness* epistemically significant and not simply *ad hoc*.

2. The Role of Possibility Spaces

In the introduction I said that I would defend two major claims: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. This section will address the first claim by giving a more detailed account of the conceptual structure of narrative explanations and why the role of possibility spaces is so central to them.

When confronted with a natural historical problem (e.g. accounting for the processes involved in the formation of atoll reefs, say (see Ghiselin 1969)) it is my claim that what we are confronted with is, in fact, a space of *possible* histories. That is, when the historical scientist attempts to answer the question, “What geological process accounts for the formation of atoll reefs?” she understands—perhaps implicitly—that there are a number of ways things *might* have gone: she sees many possible histories. This space of possible histories essentially generates a contrasting set of possible explanations, each possible history corresponding roughly to one

hypothetical solution to the problem.[2] Obviously there's just one causal history that actually obtained, but the evidential situation is such that this history is not uniquely fixed from an epistemic perspective (see Roth 2017). The historical scientist's explanatory task then consists in finding the best approximation of the true causal history.

A nice example of this sort of reasoning process can be glimpsed in the debates over speciation processes among evolutionary biologists and paleobiologists. Stephen J. Gould and Niles Eldredge (1972) developed the theory of *punctuated equilibria* to account for the pattern of speciation witnessed in the fossil record. The idea of punctuated equilibria, in brief, holds that evolutionary change occurs in sudden bursts (on geological timescales, anyway), followed by long periods of relative evolutionary stasis. The going theory of evolutionary change at the time held to *phyletic gradualism*—the idea that the pace of evolution is slow and relatively uniform (see Turner 2011). Each of these alternatives is broadly consistent with the available fossil evidence. Phyletic gradualism takes the view that the evolutionary process is gradual, and that the fossil record is very patchy. The putatively patchy character of the fossil record means that we shouldn't expect to be able to use it as a tool for faithfully reading off patterns of speciation in the actual history of life. The theory of punctuated equilibria has it that the fossil record is relatively faithful to evolutionary history, meaning that the fossil record *does* have some explanatory import with respect to uncovering important evolutionary patterns (like speciation). The evidence in the fossil record can support either interpretation.

Consider another example, this time from geology. 19th century geologists were confronted with a fascinating geological puzzle involving what were called 'erratic blocks'.

[2] I'm certainly *not* claiming that the historical scientist is in a position to generate or realize all possible histories, as the number of such alternatives is plausibly infinite. But certainly it's possible to generate quite a few, and it seems that in fact we usually do.

Seattle, WA; 1-4 November 2018 -59-

These hulking slabs of (usually) granite are found miles away from any related rocks, and so the obvious question to be answered is, "How did such a large piece of granite come to be deposited here?" In 1820s Europe the answer was not immediately obvious. One well-documented case involved a granite erratic in Switzerland, which was determined to be composed of primary rocks of Alpine origin, but resting on a limestone formation many hundreds of miles from any mountains (see Rudwick 2014, pp. 117-25). Several explanations were offered: that it was deposited by the waters of the Noahic deluge; that it was carried and deposited by waters traveling down the Alps from a broken mountain dam; and only later that it was carried by glacial ice and then deposited after a subsequent melt. The process of adjudicating between each such purportedly explanatory histories (whether evolutionary patterns or seemingly bizarre geological deposits) is the subject of the next section.

It's important to stress that the evidential underdetermination of historical hypotheses is quite different than underdetermination in science more broadly. Turner (2007) argues convincingly that the problem of underdetermination is rather severe in the historical sciences given that natural processes actively destroy the evidential traces on which historical scientists rely.[3] There are two points that make this worthy of note. First, it is precisely for this reason that the explanatory task of the historical scientist *necessarily* involves the generation of a possibility space. If we can think of a natural history as a story concerning the artifacts of the natural world, then what the world presents us with is a story that's missing a great many pages. The unfortunate fact of the matter is that there are many ways of filling those pages in, each of which

[3] Turner appeals to the role played by background theories in the historical sciences to motivate his point. Here, there relevant theory is *taphonomy*, which describes the mechanisms by which the relevant evidence is destroyed (remineralization, decomposition, etc.).

is broadly compatible with our evidential situation.[4] Second, widespread underdetermination is what motivates the earlier insight that the explanatory aspiration of historical science is to give the best *approximation* of the true causal history. It is implausible to think that any of the historical hypotheses we generate will fill in the missing pages perfectly, but we can have reasons to think that some hypotheses outperform others (of which more to come).

To summarize, possibility spaces are ineliminable from narrative explanations because of our epistemic position relative to the evidence at hand. What we want is to develop a causal history that explains the phenomenon in question (e.g. erratic blocks and evolutionary patterns), but right away we realize that many different and mutually incompatible histories could—hypothetically—do the trick. The construction of a space of live possibilities allows us to have some degree of confidence that we've explored the relevant alternatives.[5] Once we've developed a space of possibilities, the initial question (such as, "What accounts for the formation of atoll reefs?") becomes importantly *contrastive*: "Why *x* and not *x'*?" where *x* and *x'* are alternative possible causal histories accounting for the target phenomenon. We want to know how it is that possibilities come to be "foreclosed" upon as a narrative explanation develops, as Beatty (2016) puts it.

3. Causal Mechanisms and Hypothesis Adjudication

[4] See Turner (2011) chapter 2 for more in-depth discussion.

[5] There's a way of reading this that might tempt one to see this as something akin to *inference to the best explanation*. Any such connection is largely superficial. The primary reason for this is that the explanatory scheme that I'm outlining is not meant to be making any especially strong claims about the strength of an explanation as related to its connection to reality. Perhaps none of the causal histories we generate are very accurate as descriptions of the true causal history.

I now turn my attention to an explication and defense of (2): adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. Causal mechanisms are what provide reasons for preferring one possible causal history over another as regards the space of possible histories generated by the natural historical problem at hand.

3.1. Mechanistic set-ups-

Because contingency is generally seen as playing such a fundamental role in natural historical contexts, the relevant mechanisms are not likely to be cashed out in terms of 'invariances' and 'regularities,' as is common in other scientific contexts (see Havstad 2011; Darden and Craver 2002). For the purposes of natural history we might instead think in terms of a more minimal conception of causal mechanisms that I'll call *mechanistic set-ups*. A mechanistic set-up differs from paradigmatic mechanisms (as in Glennan (2002))[6] in that it will often be the case that mechanistic set-ups are the result of one-off circumstances. Paradigmatic mechanisms characterize causal systems that are largely stable across time (think of protein synthesis, for instance). Mechanistic set-ups are not stable across time in this way, but still render outcomes causally expectable given that the right antecedent conditions obtain. That is, given that the right antecedent conditions obtain (and this may, of course, be a *highly contingent* affair), the causal output of the system is fully determined—we have a case of mechanical causal output.

Nancy Cartwright and John Pemberton (2013) give a simple example of a mechanistic set-up using a toy sailboat. When the toy boat is placed in the water it displaces enough liquid to

[6] "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations."

stay afloat; it has a wind-catching device for locomotion; the wind-catching device is acted about by wind gusts in order to achieve locomotive action. If we take this example as having to do with the actions of an *agent* that brings about the mechanistic set-up then we might incline toward an interpretation of the situation in terms of paradigmatic mechanisms. But imagine there's no agent involved at all; that is, let it be the case that nobody placed the boat on the water, and likewise nobody chose any windy day in particular for the use of the boat. Instead suppose that it is a series of contingent events (a child threw the boat in the garbage, it fell out of the garbage truck on the highway, and is now on the surface of a local pond, etc.) that have made things such that the boat is at some later time moving across the top of the water in the expected way.

The one-offness of the circumstances in the revised toy boat example doesn't seem to make the situation non-mechanistic in character. Rather, the mechanism just isn't stable across time in the same way paradigmatic mechanisms are. This is a mechanism in a more minimal sense: it is a mechanistic set-up. In other words, the realization of appropriate antecedent conditions renders the outcome causally expectable, even though the antecedent conditions are highly contingent.[7]

This case is so simple that it won't have much bite against Currie. Recall that Currie's claim is that mechanisms show to be of no use in *complex* narratives. In these cases the explanatory targets are *diffuse*, meaning that they involve complex networks of causal contributors (Currie 2014). An example of a diffuse target is Sauropod gigantism, Gigantism involves, at least, skeletal pneumatization, ovipary, increased basal metabolic rate, etc. Nothing seems to unify such causal contributions, and so there is no *mechanism* for gigantism, according to Currie—the explanatory target is *too diffuse* in complex narratives.

[7] See chapter 3 of Conway Morris (2003) for an in-depth discussion.

3.2. Abiogenesis, mechanistic set-ups, and hypothesis adjudication-

Abiogenesis, I argue, qualifies as a minimal mechanistic set-up in the sense just argued for. That is, the set of facts that determined the development of the very first self-replicating, heterotrophic organisms are plausibly subject to a high degree of contingency (see Conway Morris 2003), but even so, life is a deterministic consequence of just such a contingent set of facts.[8] Further, the instances that the theory aims to explain (e.g. self-replicating molecular systems; heterotrophic metabolic systems; protective membrane enclosures, etc.) are diffuse in the same sense as Sauropod gigantism. My aim here is not to give a full theoretical survey of abiogenesis, but instead to provide just enough content to justify the claim that work in this area fulfills the description of narrative already given, and that causal mechanisms play an important explanatory role, specifically to do with hypothesis adjudication.

Probably the first serious theoretical work on the origins of life is A.I. Oparin's 1923 *The Origins of Life* (Falk and Lazcano 2012). The basic theoretical framework is familiarly Darwinian. Oparin had in mind a model of biological origins whereby life comes on-line in stages, rather than all at once. The prebiotic world, on this view, was one of something approximating 'molecular competition.' For Oparin this amounted to chemical assemblages witnessing differential stability, approximately underwriting a growth model of molecular evolution (Falk and Lazcano 2012; Pigliucci 1999). The primary thing to be explained, on this model, was the development of heterotrophic metabolism. Metabolic pathways are so complex

[8] Some recent work in origins of life research may end up giving reasons to question the assumed contingency of life's emergence. See Kauffman (1993) for a classic treatment of the "self-organization" thesis, and England (2015) for more recent theoretical developments.

that Oparin thought their development must be accounted for in a basically stepwise fashion. Differential stabilities of chemical assemblages would make it such that certain molecules would make up increasingly large proportions of the chemical 'population,' making them live candidates for further downstream innovation (like complex metabolic pathways).

Oparin-type selection models have mostly—though perhaps not entirely—fallen by the wayside. Contemporary work is focused primarily on accounting for the possibility of self-replication and autocatalysis (Penny 2005). The thought is that biological origins must be accounted for in something like a two-step process, one involving the development of self-replicating material suitable for hereditary mechanisms, and another for things like metabolism and heterocatalytic functions like protein construction (Falk and Lazcano 2012; Conway Morris 2003). One of the more promising research strains in this area concerns what's known as the 'RNA World' (Conway Morris 2003). It's widely believed to be the case that the first replicators were RNA (or RNA-like) molecules. So, RNA World researchers are attempting to simulate the conditions of the prebiotic Earth in the laboratory in order to see whether the RNA model of biological origins can carry its empirical weight.

Of note for the purposes of this paper is that the dispute between metabolism-first and replication-first models of abiogenesis is precisely over whether the causal mechanisms in play can adequately account for the target phenomenon: namely, the development of living organisms in the ancient history of Earth. H.J. Muller developed a theoretical agenda stressing the need for self-replicators at the historical foundations of life (Falk and Lazcano 2012). Oparin took heterotrophic metabolic pathways as the primary puzzle to be solved (Oparin 1938; Falk and Lazcano 2012). The replication-first view has emerged as the going view among contemporary researchers primarily because it offers a more plausible mechanism for life's early development.

In order to build complex metabolic pathway it seems like it's first necessary to have a genome space that's large enough to enable downstream innovation of complex functions. So it is that the replication-first view and the research agenda dictated by projects like RNA World are taken to be more explanatory than Oparin-type explanations given in terms of selection among molecular assemblages.

4. Putting Things Together

Let's recall once more the two key claims being advanced: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. Widespread underdetermination in the historical sciences leads to the persistent appearance of possibility spaces as specified by (1), and the development of adequate causal mechanisms specified under (2) enhances our ability to adjudicate the alternatives we're faced with. Causal mechanisms put us in a position to address the contrastive question, "Why *x* and not *x'*?" Causal mechanisms are the devices by which historical counterfactuals become foreclosed upon in the sense of Beatty (2016).

Because explanation in the historical sciences is contrastive in the above sense, I argue that some notion of mechanism is involved in *every* case of successful narrative explanation. Currie (2014) argues that causal mechanisms are appropriate only for the purposes of simple narratives apt to be embedded in terms of regularities. Complex narratives with their diffuse explanatory targets require something more piecemeal that doesn't count as a causal mechanism. My more minimal conception of causal mechanisms given in terms of *mechanistic set-ups* sheds light on why this can't be right. Mechanistic set-ups aren't stable across time like paradigmatic

mechanisms, and yet we have good reason to think that the consequences of such set-ups are mechanistically determined (see Penny 2005; Glennan 2010).[9] It is just this sort of conception of mechanism that helps us to make sense of explanatory success in abiogenesis (such as it is).

Surely the genesis of the first biotic creatures is every bit as diffuse an explanatory target as Sauropod gigantism. I've argued (and I think convincingly) that it is precisely due to the adequacy of some underlying mechanism that one explanatory agenda in abiogenesis has been accepted over the alternatives. The complexity of the narrative and the diffuseness of the explanatory target appear to be beside the point. Without an adequate mechanism—however minimally construed—we can't answer the contrastive question, and so we have no explanation at all.

5. Objection and a Reply

According to Currie (2014) mechanistic set-ups (*ephemeral mechanisms* (Glennan 2010)) look like they're simply pointing to claims about sensitivity to initial conditions. If that's right, then there's a problem, because causal processes in natural historical contexts are often thought to be contingent not just in the sense that they display sensitivity to initial conditions. Such processes are taken to be subject to contingencies in a more robust sense involving "causal cascades" themselves (Currie 2014). It is not unreasonable, for instance, to think that whether a chemical assemblage will manage to hit the right configuration and produce a self-replicating RNA strand is not just a matter of realizing the right set-up conditions (independent of the chances of hitting

[9] Penny notes some interesting experimental results in which living organisms are frozen to near absolute zero, meaning that all information concerning the positions and velocities of the particles in their make-up is lost. They can, nonetheless, be successfully reanimated. Given that the only information that's retained after such a deep freezing involves the chemical structure of the organisms, a natural inference is that 'life' is a mechanical consequence of chemical parts.

on such a configuration). Whether the chemical elements enter into the appropriate causal relations for manifesting autocatalysis might *itself* be a probabilistic matter. Having the right elements might not be all you need—you might need the right elements plus a bit of probabilistic luck. Objective probabilities of this sort might do some damage to the mechanistic account, since it would seem not to be the case that an explanandum *just follows* from a causal set-up. The force of this objection is at least partly dependent on one's answer to the question of where in the world we ought to 'place' objective chances (if there are any).

Most of our intuitions about objective probabilities (probably) derive from our ongoing observations of the world. A lot of stuff in the world *just seems* chancy. We regularly speak in terms of the "odds" or "chances" of developing cancer and the like. Simplifying quite a bit, when we say that there's a 40 percent chance that Susan will live for more than 5 years after being diagnosed with some cancer that has developed to some particular stage, what we're saying is that approximately 40 percent of people that present as cases sufficiently similar to Susan have lived for 5 years or more. One way to read this is in terms of causal indeterminacy. That is, there is really no matter of the fact at time *t* as to what will be the case at time *t'*, aside from the probabilistic facts about cancer populations. The future is (to some degree) causally open, as the causal cascades are operating in a fundamentally probabilistic way.

Such a reading, however, is by no means forced. Bruce Glymour (1998) offers a picture wherein objective probabilities are placed at the level of causal *interactions*. That is, entities *e* and *e** enter into causal interactions with each other on a probabilistic basis, but when they do, the downstream effects unfold in a fully deterministic fashion. Probabilistic partitions of the world, then, are just reflections of whether certain causal interactions became manifest in certain subpopulations or not. If 40 percent of patients with a certain cancer at a particular stage will

survive for more than five year, it's because free radicals (probabilistically) failed to enter into certain causal interactions with healthy cells. The opposite is the case for the contrasting class of fatal cases. On this picture, determinism of the relevant kind seems to be preserved. In such cases as the right causal interactions are realized, downstream effects unfold in mechanical fashion.

6. Conclusion

In this paper I argued for two main claims: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. The reason that narrative explanations involve possibility spaces has to do with our epistemic position relative to the available evidence. Undetermination so permeates the historical sciences that any problem for which we seek an explanation will involve an array of possible alternative causal histories, each of which is broadly consistent with the available evidence. It is the introduction of an adequate causal mechanism that puts us in a position to improve our epistemic lot—with a good mechanism in hand, we can begin to foreclose upon alternatives.

References

Beatty, John. 2016. “What Are Narratives Good For?” *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 58. Elsevier Ltd: 33–40. doi:10.1016/j.shpsc.2015.12.016.

———. 2017. “Narrative Possibility and Narrative Explanation.” *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 1–14. doi:10.1016/j.shpsa.2017.03.001.

Cartwright, Nancy, and John Pemberton. 2013. "Aristotelian Powers: Without Them, What Would Science Do?" in Groff & Greco (Eds.), *Powers and Capacities in Philosophy: The New Aristotelianism*. New York: Routledge.

Conway Morris, Simon. 2003. *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge: Cambridge University Press.

Currie, Adrian Mitchell. 2014. “Narratives, Mechanisms and Progress in Historical Science.” *Synthese* 191 (6): 1163–83. doi:10.1007/s11229-013-0317-x.

Darden, Lindley, and Carl Craver. 2002. "Strategies in the interfield discovery of the mechanism of protein synthesis." *Studies in History and Philosophy of Biological and Biomedical Sciences* 33: 1-28.

Eldredge, Niles, and Stephen J. Gould. 1972. "Punctuated equilibria: an alternative to phyletic gradualism," in Schopf (Ed.), *Models in Paleobiology*. San Francisco: Freeman Cooper.

England, Jeremy. 2015. "Dissipative Adaptation in Self-Driven Assembly." *Nature Nanotechnology*, 10: 919-923.

Ereshefsky, Marc. 1992. "The Historical Nature of Evolutionary Theory." In *History and Evolution*, ed. Matthew Nitecki and Doris Nitecki. New York: The SUNY Press.

Falk, Raphael, and Antonio Lazcano. 2012. “The Forgotten Dispute: A.I. Oparin and H.J. Muller on the Origin of Life.” *History and Philosophy of the Life Sciences* 34 (3): 373–90.

Ghiselin, Michael T. 1969. *The Triumph of the Darwinian Method*. Chicago: Chicago University Press.

Glennan, Stuart. 1996. “Mechanisms and the Nature of Causation.” *Erkenntnis* 44 (1): 49–71.

———. 2002. “Rethinking Mechanistic Explanation.” *Philosophy of Science* 69 (S3): S342–53.

———. 2010. “Ephemeral Mechanisms and Historical Explanation.” *Erkenntnis* 72 (2): 251–66. doi:10.1007/s10670-009-9203-9.

Glymour, Bruce. 1998. “Contrastive, Non-Probabilistic Statistical Explanations.” *Philosophy of Science* 65 (3): 448–71.

Gordon, Malcolm and Everett Olson. 1994. *Invasions of the Land*. New York: Columbia University Press.

Haldane, J.B.S. 1954. "The origin of life." *New Biology* 16: 12-27.

Havstad, Joyce C. 2011. “Problems for Natural Selection as a Mechanism.” *Philosophy of Science* 78 (3): 512–23. doi:10.1086/660734.

Hull, David. 1975. “Central Subjects and Historical Narratives.” *History and Theory* 14 (3): 253–74.

Jeffares, Ben. 2008. “Testing Times: Regularities in the Historical Sciences.” *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 39 (4). Elsevier Ltd: 469–75. doi:10.1016/j.shpsc.2008.09.003.

Kauffman, Stewart. 1993. *The Origins of Order: Self Organization and Selection in Evolution*. Oxford: Oxford University Press.

Mink, Louis O. 1970. "History and Fiction as Modes of Comprehension." *New Literary History*, 1 (3): 541-558.

Oparin, A.I. 1938. *The Origin of Life*. New York: MacMillan.

Penny, David. 2005. “An Interpretive Review of the Origin of Life Research.” *Biology & Philosophy* 20 (4): 633–71. doi:10.1007/s10539-004-7342-6.

Pigliucci, Massimo. 1999. "Where do we come from? A humbling look at the biology of life's origin." *Skeptical Inquirer* 99: 193-206.

Ricoeur, Paul. 1984. *Time and Narrative (Volume 1)*. Chicago: University of Chicago Press.

Roth, Paul A. 2017. “Essentially Narrative Explanations.” *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 1–9. doi:10.1016/j.shpsa.2017.03.008.

Rudwick, M.J.S. 2014. *Earth's Deep History: How It Was Discovered and Why It Matters*. Chicago: Chicago University Press.

Sepkosi, David. 2012. *Rereading the Fossil Record: The Growth of Paleontology as an Evolutionary Discipline*. Chicago: Chicago University Press.

Sunstein, Cass R. 2016. “Historical Explanations Always Involve Counterfactual History.” *Journal of the Philosophy of History* 10 (3): 433–40. doi:10.1163/18722636-12341345.

Turner, Derek. 2013. “Historical Geology: Methodology and Metaphysics.” *Geological Society of America Special Papers* 502 (2): 11–18. doi:10.1130/2013.2502(02).

———. 2007. *Making Prehistory: Historical Science and the Scientific Realism Debate*. Cambridge: Cambridge University Press.

———. 2011. *Paleontology: A Philosophical Introduction*. Cambridge: Cambridge University Press.

A Better Foundation for Public Trust in Science

S. Andrew Schroeder
Claremont McKenna College/Princeton University
aschroeder@cmc.edu

draft of 15 June 2018

Abstract. There is a growing consensus among philosophers of science that core parts of the scientific process involve non-epistemic values. This undermines the traditional foundation for public trust in science. In this paper I consider two proposals for justifying public trust in value-laden science. According to the first, scientists can promote trust by being transparent about their value choices. On the second, trust requires that the values of a scientist align with the values of an individual member of the public. I argue that neither of these proposals work and suggest an alternative that does better: when scientists must appeal to values in the course of their research, they should appeal to *democratic values*, the values of the public or its representatives.

1. Introduction

The American public's trust in science is a complicated matter. Surveys reveal that trust in science has remained consistently high for decades, and scientists remain among the most highly-trusted professional groups (Funk 2017). However, within some segments of society (especially conservatives) trust has declined significantly (Gauchat 2012), and there are obviously serious gaps in trust on certain issues, such as climate change, vaccine safety, and GM foods (Funk 2017). The picture, then, is a complex one, but on balance it is clear that things would be better if the public placed greater trust in science and scientists, at least on certain issues.

As a philosopher, I am not in a position to determine what explains the lack of trust in science, nor to weigh on what will in fact increase trust. Instead, in this paper I will look at the question of what scientists can do to *merit* the public's trust — under what conditions the public *should* trust scientists. Indeed, it seems to me that we need to answer the normative question first: if we take steps to increase

public trust in science, our goal should not simply be to make scientists *trusted*, we should also want them to be *trustworthy*.

In what follows, I'll first explain how recent work in the philosophy of science undermines the traditional justification given to the public for trusting science. I'll then consider two proposals that have been offered to ground public trust in science: one calling for transparency about values, the second calling for an alignment of values. I'll argue that the first proposal backfires — it rationally should *decrease* trust in science — and the second is impractical. I'll then present an alternative that is imperfect, but better than the alternatives: when scientists must appeal to values in the course of their work, they should appeal to *democratic values* — roughly, the values of the public or its representatives.

2. Trust and the Value-Free Ideal

Why should the public trust scientists? The typical answer to that question points to the nature of science. Science, it is said, is about facts, and not values. It delivers us objective, verifiable truths about the world — truths not colored by political beliefs, personal values, or wishful thinking. Of course, there are scientists who inadvertently or intentionally allow ideology to influence their results. But these are instances of *bad* science. Just as we should not allow the existence of incompetent or corrupt carpenters to undermine our trust in carpentry, we should not allow the existence of incompetent or corrupt scientists to undermine our trust in science. So long as we have institutions in place to credential good scientists and root out corrupt ones, we should trust the conclusions of science.

There is, unfortunately, one problem with this story: science isn't actually like that. In the past few decades, philosophers of science have shown that even good science requires non-epistemic value judgments. Without wading into the nuanced differences between views, I think it is fair to say that there is a consensus among philosophers of science that non-epistemic values can appropriately play a role in at

least some of the following choices: selecting scientific models, evaluating evidence, structuring quantitative measures, defining concepts, and preparing information for presentation to non-experts.[1]

These value choices can have a significant impact on the outcome of scientific studies. Consider, for example, the influential Global Burden of Disease Study (GBD). In its first major release it described itself as aiming to "decouple epidemiological assessment from advocacy" (Murray and Lopez 1996, 247). In the summary of their ten volume report, the authors describe their study as making "a number of startling individual observations" about global health, the first of which was that, "[t]he burdens of mental illnesses…have been seriously underestimated by traditional approaches… [P]sychiatric conditions are responsible…for almost 11 per cent of disease burden worldwide" (Murray and Lopez 1996, 3). Many others have cited and relied on the GBD's conclusions concerning the magnitude of mental illness globally (Prince *et al*. 2007). And nearly two decades later, the same GBD authors, in commenting on the legacy of the 1996 study, proudly noted that it "brought global, regional, and local attention to the burden of mental health" (Murray *et al*. 2012, 3).

It turns out, however, that the reported burden of mental health was driven largely by two value choices: the choice to "discount" and to "age-weight" the health losses measured by the study. Discounting is the standard economic practice of counting benefits farther in the future as being of lesser value compared to otherwise similar benefits in the present, and age-weighting involves giving health losses in the middle years of life greater weight than otherwise similar health losses among infants or the elderly. Further details about discounting and age-weighting aren't relevant to this paper; all we need to note is that the study authors acknowledged that each reflects value judgments, and that a reasonable case could be made to omit them (Murray 1996; Murray *et al*. 2012).[2] Given other methodological choices made by the authors, these two weighting functions combine to give relatively more weight to health

[1] On these points see e.g. Reiss (2017) and Elliott (2011).

[2] Indeed, in 2012 the GBD ceased age-weighting and discounting. There was also a third value choice that drove the large burden attributed to mental health: the choice to attribute all suicides to depression (Murray and Lopez 1996, 250). Because I do not know precisely how this affected the results, I set it aside here. For much more on discounting, age-weighting, and other value choices in the GBD, see Schroeder (2017).

conditions which (1) commonly affect adults or older children (rather than the elderly or young children), (2) have disability (rather than death) as their primary impact, and (3) have their negative effects relatively close to the onset or diagnosis of the condition (rather than far in the future). It should not be surprising, then, that when the GBD authors ran a sensitivity analysis to see how the decision to discount and age-weight affected the results, they discovered that the conditions most affected by these choices — unipolar major depression, anaemia, alcohol use, bipolar disorder, obsessive-compulsive disorder, chlamydia, drug use, panic disorder, post-traumatic stress disorder — were largely composed of mental health conditions (Murray and Lopez 1996, 282). Overall, the global burden of disease attributable to psychiatric conditions drops from 10.5% to 5.6%, when the results are not age-weighted or discounted (Murray and Lopez 1996, 261, 281).

I don't want to comment here on the wisdom of the GBD scientists' decision to discount and age-weight.[3] They offer clear arguments in favor of doing so and many other studies have done the same, so at minimum I think their choices were defensible. The point is that what was arguably the top-billed result of a major study — a result which was picked up on by many others, and which was still being proudly touted by the study authors years later — was not directly implied by the underlying facts. It was driven by a pair of value judgments. Had the GBD scientists had different views on the values connected to discounting and age-weighting, they would have reported very different conclusions concerning the global impact of mental illness.[4]

This case is not unique. The dramatically different assessments given by Stern and Nordhaus on the urgency of acting to address climate change can largely be traced to the way each valued the present versus the future (Weisbach and Sunstein 2009). Similar conclusions are plausible concerning the value choices involved in classifying instances of sexual misbehavior in research on sexual assault, the value

[3] I do so in Schroeder (unpublished-a).

[4] Although the sensitivity analysis was conducted by the original study authors, they do not draw any connection to their prominent claims concerning the global extent of mental illness. To my knowledge, this paper is the first to do so.

choices impacting the modeling of low-level exposures to toxins (Elliott 2011), and the value choices involved in constructing price indices (Reiss 2008).

A natural — and not implausible — response to these cases is to suggest they are outliers. Although some scientific conclusions are sensitive to value choices, the vast majority are not. The Earth really is getting warmer and sea levels really are rising, due to human activity. Vaccines really do prevent measles and really don't cause autism. These conclusions are not sensitive in any reasonable way to non-epistemic value judgments made by scientists in the course of their research. The problem, however, is that there is no clear way for a non-expert to verify this — to tell which cases are the outliers and which are not. This, I think, justifies a certain amount of skepticism. "Although some of our conclusions do depend on value judgments, trust us that *this* one doesn't," isn't nearly as confidence-inspiring as, "Our conclusions depend only on facts, not values."

I conclude, then, that rejecting the view of science as value-free, combined with high-profile examples of scientific conclusions that do crucially depend on value judgments, undermines the claim of science to public trust in a significant way. In other words, it explains why it may be rational for the public to place less trust in the conclusions of science on a broad range of issues — including in areas, such as climate change and vaccine safety, where major conclusions are not in fact sensitive to different value judgments.[5]

3. Grounding Trust in Transparency

Good science is not value-free, which undermines the standard justification given for trust in science. What, then, can scientists do to merit the public's trust? The standard response has been to appeal to transparency. If values cannot or should not be eliminated from the scientific process, scientists

[5] For similar conclusions see Douglas (2017); Wilholt (2013); Irzik and Kurtulmus (*forthcoming*); and Elliott and Resnik (2014).

should be "as transparent as possible about the ways in which interests and values may influence their work" (Elliott and Resnik 2014, 649; *cf.* Ashford 1998; Douglas 2008; McKaughan and Elliott 2018). Obviously, in order for this proposal to work, scientists would need to be aware — much more aware than most are today — of the ways in which value judgments influence their work. But, since we have independent reason to want such awareness, let us assume that calls for transparency are accompanied by a mechanism for increasing such awareness by scientists.

Would such a proposal work? Transparency about values can help ground trust in some situations, but I see no reason to think that it should broadly support public trust in science. Transparency is only useful in supporting — as opposed to eroding — trust if it enables the recipient of that information to determine how it has affected the author's conclusions. (Knowing I have a conflict of interest will typically reduce your trust in what I tell you, unless you can determine how that conflict influenced my conclusions.) Transparency, then, will only promote trust in a robust way if the public understands how value choice influenced the results, and understands what alternative value choices could have been made and how they would have influenced the results. These criteria may be satisfiable when the effect of a value choice is relatively simple. Suppose, for example, that a scientist classifies non-consensual kissing as "sexual assault", rather than "sexual misconduct", on the grounds that she believes it has more in common with rape (a clear instance of sexual assault) than it does with contributing to a sexualized workplace (a clear instance of sexual misconduct). The value judgment here is relatively simple to explain, an alternative classification is obvious, and (if the statistics involved are simple) the effect of alternative classification on the study may be relatively straightforward. So transparency could work here.

Many value choices, however, are much more complex. Think about choices embedded in complex statistical calculations — for example, those involved in aggregating climate models (Winsberg 2012) or in calculating price indices (Reiss 2008). In cases like these, it will be very hard to clearly explain the importance of any individual value choice and harder still to explain what alternative choices

could have been made. Further, many studies involve a large number of value judgments. Schroeder (2017), for example, identifies more than ten value choices which non-trivially influenced the Global Burden of Disease Study's results. Even if each of those value choices could be explained individually, it would be virtually impossible for a non-expert to figure out the interaction effects between them.

What these cases show is that even if scientists make a serious effort at transparency — not simply listing their value judgments, but attempting to explain how those judgments have influenced their results — in many cases it simply won't be possible to communicate to the public how those values have impacted their work.[6] And, if the public can't trace the impact of those values, transparency doesn't amount to much more than a warning — a reason to *distrust*, rather than to trust. A parallel realization can be seen in the way many medical schools and journals have handled researchers' conflicts of interest. Whereas in the past disclosures of conflicts of interest — essentially, transparency — were often regarded as sufficient; many have now realized that merely knowing about such conflicts does not appreciably help a reader to interpret a study. There is thus a growing move towards banning all significant conflicts of interest.[7]

4. Grounding Trust in an Alignment of Values

The previous section argued that transparency about values is not typically a solution to the problem of public trust in science. That problem, we can now see, was not caused by the fact that values were *hidden*; it was caused by the fact that the values of scientists may *diverge* from the values of any

[6] McKaughan and Elliott (2018, and in other works) suggest that scientists, through a particular sort of transparency, seek to promote "backtracking" — that is, to enable non-experts to understand how values have influenced scientists' results and to see how those results might have looked given alternative values. They seem to suggest that, at least in the cases they consider, this will frequently be possible. I am claiming that this will not generally be feasible. See Schroeder (unpublished-a) for a more detailed discussion of a particular case.

[7] See e.g. <https://ari.hms.harvard.edu/interim-policy-statement-conflicts-interest-and-commitment>

individual member of the public.[8] To promote public trust in science, then, it seems that we need to eliminate that divergence. This is the insight that motivates Irzik and Kurtulmus (*forthcoming*; *cf*. Douglas 2017; Wilholt 2013), who argue that what they call "enhanced" trust requires that a member of the public knows that a scientist has worked from value choices that are in line with her own.

If this proposal were feasible, I think it would provide a good foundation for trust. And, in certain limited cases, it may be feasible. When science is conducted by explicitly ideological organizations, members of the public may be able to make quick and generally accurate judgments about what values scientists hold, and accordingly may be able to seek out research done by scientists who share their values. (A pragmatic environmentalist, for example, might be confident that scientists employed by the Environmental Defense Fund are likely to share her values.)

Most science, however, is not conducted by explicitly ideological organizations. In these cases, it will typically be very hard for members of the public to confidently determine whether a given study relied on value judgments similar to her own. Even when this can be done (perhaps as a result of admirable transparency and clarity on the part of a scientist), it will require sustained and detailed engagement from the public, who will have to pay close attention not just to the conclusions of scientific studies, but also to their methodology. Although such close attention to the details of science would be beneficial for a great many reasons, it unfortunately is not realistic on a broad scale. There are simply too many scientific studies out there that are potentially relevant to an individual's decisions for even attentive members of the public to keep up. If our model for trust in science requires an alignment of values between the scientist and individual members of the public, trust in science can't be a broad phenomenon. Further, I don't think we want our foundation for trust in science to make that trust accessible only to those with the education and time to invest in exploring the details of individual scientific studies.

[8] It seems relevant to note here that distrust in science is greatest among those who identify as politically conservative, while studies show that university scientists in the U.S. overwhelming support liberal candidates for political office. Whether or not this in fact explains the distrust conservatives have in science, the argument thus far shows why such distrust could have a rational foundation.

Seattle, WA; 1-4 November 2018 -81-

I also — somewhat speculatively — worry that adopting this proposal would exacerbate another problem. Suppose the proposal works and, at least on some issues, members of the public are able to identify and rely upon science conducted in accordance with their own values. This, I think, might lead to a further "politicization" of science, as each side on some issue seeks scientists who share their values. Of course, once we allow a role for values in science, value-based scientific disagreement isn't necessarily a problem. Faced, for example, with one experimental design that is more prone to false positives and another that is more prone to false negatives, either choice may be scientifically legitimate. It may therefore be appropriate for more environmentally-minded citizens to rely on different studies than citizens more concerned about economic development. I worry, though, that in a culture where the public specifically seeks science done by those who share their values, it will be too easy to write off any differences in conclusions as due to value judgments — too easy for environmentalists to assume that any time pro-environment and pro-industry scientists reach different conclusions, it must be due to different underlying, legitimate value judgments. In reality, though, most such disagreements are the result of *bad* science. The worry, then, is that if we grow too comfortable with each side of an issue having its own science, it will be harder to distinguish scientific disagreements that can be traced to legitimate value judgments, from disagreements that are based on illegitimate value judgments or simple scientific error. This would be a major loss.

5. Grounding Trust in Democratic Values

I've argued that neither transparency about values nor an alignment of values can provide a broad foundation for public trust in science. Let me, then, suggest a proposal that, though imperfect, can do better. From what's been said so far, we can note a few features that a better solution should have. First, both the transparency and aligned values proposals ran into trouble because they require a great deal of attention and sophistication from the public. Most individuals simply don't have the training to

understand more technical value choices, or value choices embedded within complex calculations. And, even when such understanding is possible, it will often require a level of attention that will in practice be accessible only to the well-off. We should therefore look for a foundation for public trust which doesn't require such detailed understanding of or close attention to individual scientific studies. Second, I suggested that the aligned values proposal, in telling individuals to seek out studies conducted in accordance with their own values, could reinforce a kind of politicization that may have bad consequences. It would be better to find a proposal that wouldn't so easily divide scientists and the public along ideological lines. Third, the problem with the transparency proposal (which the aligned values proposal tried, impractically, to address) was that values, even if transparent, can be alien. In order for an individual to truly trust science, that science must be built on values that have some kind of legitimacy for her.

I think scientists can satisfy two-and-a-half of these three criteria by appealing to *democratic values* — the values of the public and its representatives — when value judgments are called for in the scientific process. The details of this proposal go beyond what I can say here.[9] But, briefly, the idea is that we look to political philosophy to tell us how to determine the (legitimate) values representative of some population. In some cases, those values might be the output of a procedure, such as a deliberative democracy exercise, a citizen science initiative, or a public referendum.[10] In other cases, it might be more appropriate to equate a population's values with the views, suitably "filtered" and "laundered", currently held by its members. ("Filtering" may be necessary to remove politically illegitimate values, e.g. racist values, and "laundering" to clean up values that are unrefined or based on false empirical beliefs.) In cases where there is a broad social consensus, that might count as the relevant democratic value; in cases where there is a bimodal distribution of values, we might say that there are two democratic values; etc.

---

[9] See Schroeder (unpublished-b) for a bit more. Many other philosophers have argued that there should be an important place for democratic values in science. See, for example, Kitcher (2011), Intemann (2015), and Douglas (2005).

[10] The extensive literature on "mini-publics" offers a promising starting point. See e.g. Escobar and Elstub (2017).

Seattle, WA; 1-4 November 2018 -83-

Suppose, then, that political philosophers, informed by empirical research, can give us a way of determining democratic values. I suggest that when value judgments are called for within the scientific process,[11] scientists should use democratic values when arriving at their primary or top-line results — the sort of results reported in an abstract, executive summary, or in the initial portions of the analysis. Scientists could then offer a clearly-designated alternative analysis based on another set of values, e.g. their own. I think this proposal can address two of the concerns with which I began this section, and can make some progress towards answering the third.

Let us first consider the too-much-attention and politicization problems. On the democratic values proposal, if an individual can trust that a study was competently carried out — a matter I'll return to below — then she can know, without digging into the methodological details, that its conclusions are based on objective facts plus democratic values.[12] This means that, in most cases, the public need not pay detailed attention to the methodological details of individual studies — thus solving the too-much-attention problem. Further, if scientific conclusions are based on objective facts plus democratic values, any two scientists investigating the same problem in the same social and political context should reach roughly the same conclusion. This recovers a kind of objectivity for science — not objectivity as freedom from values, but objectivity as freedom from personal biases. On this picture, the individual characteristics of a scientist should have no impact on her conclusions — a conception of objectivity that has been defended on independent grounds (Reiss and Sprenger 2014; *cf.* Daston and Galison 2007 on "mechanical objectivity"). If they are both doing good science, the environmentalist and the industrialist should reach the same top-line conclusions. And if the environmentalist and industrialist reach different

[11] This proposal is restricted to value judgments that arise within the scientific process. In particular, I do not mean for it to apply to problem selection. Scientists should be free to choose research projects that are not the projects that would be chosen by the general public. (The public, however, is under no obligation to fund such projects.) I treat the choice of research topics differently than choices that arise within the course of research because I think that scientists have different rights at stake in each case. For some related ideas, see Schroeder (2017b).

[12] There may also, of course, be methodological choices not based on non-epistemic values (including choices based on epistemic values). I set these aside here, since the problems of trust I'm concerned with don't arise in the same way from them.

top-line conclusions, it means that one or the other has made some sort of error. This, I think, provides a solution to the politicization problem: on the democratic values proposal, good science (at least in its primary analyses) will speak with a single voice.

The democratic values proposal therefore solves two of the three problems we noted above. Of course, it only does so if the public can be confident that scientists really are making use of democratic values. Why should the public assume that? Right now, I think the answer is: they shouldn't! For the democratic values proposal to work, it must be accepted by a significant portion of the scientific community, or by an easily-identifiable subset of the scientific community. If that were to happen, though, then the problem here becomes the more general one of how the public can trust scientists to enforce their own norms. The procedures and policies now in place work reasonably well, I think, to expose unethical treatment of research subjects, falsification of data, and certain other types of misconduct. I am therefore optimistic that, given a greater awareness of the role value judgments play in scientific research, a system could be devised to identify scientists who depart from a professional norm requiring the use of democratic values.

6. Science, Values, and Democracy

I've argued that the democratic values proposal can address two of the problems that faced the alternative views. But what about the third? On the transparency proposal, the values of scientists can truly be alien. If a scientist conducts research based on her own values, then, unless I happen to share those values, I have no meaningful relationship to those values. If, however, a scientists appeals to democratic values, then there is a relationship, even if I don't share those values. If democratic procedures or methods were carried out properly, then my values were an input into the process which yielded democratic values. My values are, in a sense, represented in the output of that process. This, in turn, means that those values should have a kind of legitimacy for me. In a democracy, we regularly

impose non-preferred outcomes on people when they are out-voted. So long as democratic procedures are carried out properly, this seems to be legitimate — not ideal, perhaps, but better than any available alternative. On the democratic values proposal, then, when a particular scientific conclusion is uncontested, the public can trust that that conclusion is one drawn solely from the facts, plus perhaps the values that *we* share. For most of us, who don't have the time, inclination, or ability to dig into the details of each scientific study we rely on, or who have a strong commitment to democracy, that will be enough.

I think that the foregoing provides a reasonable answer to the alien values concern. It is of course not a perfect answer. It would be better, at least from the perspective of trust, to get each member of the public access to "personalized" science conducted in accordance with her values. This, however, is impractical, as we saw when discussing the aligned values proposal. So long as that is the case, there is no way to accommodate everyone. Democratic values seem like a reasonable compromise in such a situation.

All of that said, it would be nice if we could say a bit more to those ill-served by democratic values. What should we say, for example, to an individual who knows that her values lie outside the political mainstream on some issue and is therefore distrustful of science done with democratic values on that issue? The first thing to note is that, in such cases, the democratic values proposal fares no worse (or at least not much worse) than the transparency or aligned values proposals. The democratic values proposal is fully consistent with transparency - something we have independent reason to want. So, in cases where the transparency proposal works (e.g. cases where the value choices are few, easy to understand, and computationally simple), the same advantages can be had with the democratic values proposal. Individuals who disagree with a particular value judgment and have the time and expertise to do so can determine how results would have looked under a different set of value judgments. Also, recall that I am proposing only that primary or top-line results be based on democratic values. In cases where value judgments can make a big difference — as in the Global Burden of Disease Study case discussed earlier — we might hope that scientists who hold contrary values will note the dependence of those

results on values by offering secondary, alternative analyses that begin from different value judgments. Those who have the time and expertise to dig into the methodology of scientific reports can do so, seeking out results based on values they share, as the aligned values proposal would recommend.

If the foregoing is correct, the democratic values proposal does better than the alternatives in most cases, and no worse in others. That should be sufficient reason to prefer it. But I think we can say a bit more. In what cases is the complaint from minority values most compelling? It is not, I think, when it comes from people whose values lie outside the mainstream on some issues, but within the mainstream on many other issues. The much more compelling complaint comes from people whose values consistently lie outside the mainstream — people who are consistently out-voted. Oftentimes (though of course not always) when this happens, it involves individuals who are members of groups that are or have been marginalized by mainstream society. Think, for example, of cultural or (dis)ability-based groups whose values and ways of life have been consistently treated as being less valuable and worthy of respect than the values and ways of life of the majority.

I think the democratic values proposal has two important features that can partially address such complaints. First, remember that the democratic values proposal launders and filters the actual values held by the public. Certain values — e.g. racist or sexist ones — conflict with basic democratic principles of equal worth, and so cannot be candidate democratic values. Thus, even in a racist society, telling scientists to work from democratic values will not tell them to work from racist values.[13] Second, in what I regard as its most plausible forms, democracy is not a form of government based on one person-one vote. It is a form of government based on the idea that all citizens are of equal worth and have a right to equal consideration. This suggests that, in cases where minority values are held by a group that is or has been the subject of exclusion or discrimination, democratic principles may sometimes require giving those values extra weight, or a voice disproportionate to their statistical representation in the population, as a way of accounting or compensating for past unjust treatment. Thus, democratic principles may in

[13] See Schroeder (unpublished-b) for more on this.

Seattle, WA; 1-4 November 2018 -87-

some cases require treating the values held by an excluded minority as democratically on a par with the conflicting values held by the majority.[14]

These considerations, I think, lessen the force of the complaint from minority values, especially in its most serious incarnation. But I don't think they eliminate it. There will still be people whose values will consistently be marginalized by the democratic view. In such cases, the main recourse available is an appeal to alternate results. If individuals with minority views can count on there being scientists who share those views, they can expect that the kind of alternative analysis they would prefer will be out there, at least in cases where it makes a difference. Of course, scientists are currently a rather homogeneous bunch along many dimensions. So this suggests that the call to work from democratic values provides (yet further) support for the importance of increasing diversity within the scientific community.[15]

[14] See Kelman (2000) for an example of this sort of argument in the context of disability.

[15] ACKNOWLEDGEMENTS TO BE ADDED

References

Ashford, Nicholas. 1988. "Science and Values in the Regulatory Process." *Statistical Science* 3.

Daston, Lorraine and Peter Galison. 2007. *Objectivity*. MIT Press.

Douglas, Heather. 2017. "Science, Values, and Citizens." In *Eppur si muove: Doing History and Philosophy of Science with Peter Machamer,* ed. Adams, Biener, Feest, and Sullivan. Dordrecht: Springer.

—. 2008. "The Role of Values in Expert Reasoning." *Public Affairs Quarterly* 22.

—. 2005. "Inserting the Public into Science." In *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*, ed. Maasen and Weingart. Dordrecht: Springer.

Elliott, Kevin. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. Oxford: Oxford University Press.

Elliott, Kevin and David Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives* 122.

Escobar, Oliver and Stephen Elstub. 2017. "Forms of Mini-Publics: an Introduction to Deliberative Innovations in Democratic Practice," NewDemocracy Research and Development Note, available at <https://www.newdemocracy.com.au/research/research-notes/399-forms-of-mini-publics>.

Funk, Cary. 2017. "Real Numbers: Mixed Messages about Public Trust in Science." *Issues in Science and Technology* 34.

Gauchat, Gordon. 2012. "Politicization of Science in the Public Sphere: A Study of Public Trust in the United States, 1974 to 2010." *American Sociological Review* 77.

Intemann, Kristin. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5.

Irzik, Gürol and Faik Kurtulmus. *Forthcoming*. "What is Epistemic Public Trust in Science?" *British Journal for Philosophy of Science*.

Kelman, Mark. 2000. "Does Disability Status Matter?" In *Americans with Disabilities: Exploring Implications of the Law for Individuals and Institutions,* eds. Francis and Silvers. Routledge

Kitcher, Philip. 2011. *Science in a Democratic Society*. Amherst, NY: Prometheus.

McKaughan, Daniel and Kevin Elliott. 2018. "Just the Facts or Expert Opinion? The Backtracking Approach to Socially Responsible Science Communication," in *Ethics and Practice in Science Communication* (eds. Priest, Goodwin, and Dahlstrom). Chicago: University of Chicago Press.

Murray, Christopher. 1996. "Rethinking DALYs." In *The Global Burden of Disease,* ed. Murray and Lopez.

Murray, Christopher and Alan Lopez (Eds). 1996. *The Global Burden of Disease*. Harvard University Press.

Murray, Christopher *et al*. 2012. Supplementary appendix to "GBD 2010: design, definitions, and metrics." *Lancet* 380.

Prince, Martin *et al*. 2007. "No health without mental health." *Lancet* 370.

Reiss, Julian. 2017. "Fact-value entanglement in positive economics." *Journal of Economic Methodology* 24.

—. 2008. *Error in Economics: The Methodology of Evidence-Based Economics*. London: Routledge.

Reiss, Julian and Jan Sprenger. 2014. "Scientific Objectivity." In *The Stanford Encyclopedia of Philosophy* (Winter 2017 edition), ed. Zalta.

Schroeder, S. Andrew. 2017. "Value Choices in Summary Measures of Population Health." *Public Health Ethics* 10.

—. 2017b. "Using Democratic Values in Science: an Objection and (Partial) Response," *Philosophy of Science* 84.

—. Unpublished-a. "Which Values Should We Build Into Economic Measures?" *Under review.*

—. Unpublished-b. "Communicating Scientific Results to Policy-makers," *manuscript on file with author.*

Weisbach, David and Cass Sunstein. 2009. "Climate Change and Discounting the Future: A Guide for the Perplexed," *Yale Law and Policy Review* 27.

Wilholt, Torsten. 2013. "Epistemic Trust in Science." *British Journal for Philosophy of Science* 64.

Winsberg, Eric. 2012. "Values and Uncertainties in the Predictions of Global Climate Models." *Kennedy Institute of Ethics Journal* 22.

PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association

Inferential power, formalisms, and scientific models

Ardourel Vincent*, Anouk Barberousse†, Cyrille Imbert§

* IHPST — CNRS, Université Paris 1 Panthéon-Sorbonne

† SND — CNRS, Sorbonne Université

§ Archives Poincaré — CNRS, Université de Lorraine

Abstract

Scientific models need to be investigated if they are to provide valuable information about the systems they represent. Surprisingly, the epistemological question of what enables this investigation has hardly been investigated. Even authors who consider the inferential role of models as central, like Hughes (1997) or Bueno and Colyvan (2011), content themselves with claiming that models contain *mathematical resources* that provide *inferential power*. We claim that these notions require further analysis and argue that mathematical formalisms contribute to this inferential role. We characterize formalisms, illustrate how they extend our mathematical resources, and highlight how distinct formalisms offer various inferential affordances.

1. Introduction. When analyzing scientific representations, philosophers of science are keen on mentioning that some models provide scientists with "mathematical resources" and "inferential power", but they seldom give a detailed analysis of these notions. This paper is devoted to the discussion of what appears to us as major mathematical resources, namely, formalisms. We thus present an analysis of the notion of formalism as well as examples from which we argue that formalisms should be acknowledged as major units of scientific activity.

We proceed as follows. In Section 2, we briefly review what philosophers of science have to say about mathematical resource and inferential power and observe that it is disappointing. In order to fill the gap we have identified, we put forward in Section 3 the three components we identify within the notion of mathematical resource. Section 4 is devoted to one of these components, namely, formalism. At last, in Section 5, we provide the reader with examples of how the choice of a formalism influences the type of knowledge scientists may draw from their representations.

2. Scientific representations and inferences therefrom. At what conditions can scientific models be used to gain information about target systems? First, a suitable semantic relation between the model and the system(s) that it stands for should obtain, so that by investigating the model, we can make legitimate inferences about its target system(s). This cannot be done unless nontrivial inferences about the model itself, as a mathematical object, can be carried out. Models are usually referred to by proper names (like "Ising model" or "Lotka-Volterra" model") or by expressions that highlight some of their mathematical properties (like "the harmonic oscillator" or "the ideal gas"). There is however more to be learnt about them than their prima facie properties. For example, solving the Ising model reveals more about Ising-like systems than their description as "sets of discrete variables representing magnetic dipole moments of atomic spins that can be in one of two states"; similarly, the mathematical content of an harmonic oscillator goes beyond "being a system that, when displaced from its equilibrium position, experiences a restoring force that is proportional to the displacement". Philosophers of science are aware of the need to investigate the epistemology of models and how we find out about concealed truths about model systems (Frigg,

2010, 257) but are surprisingly silent about how it is actually performed.[1] They are content with saying that the model is "manipulated" (Morgan and Morrison, 1997, chapter 2, passim) or that we can "play" with it (Hughes, 2010, 49), which are suggestive, but metaphoric characterizations.

Surprisingly, even accounts of applied mathematics and scientific representation that give central stage to their inferential role hardly analyze how it is fulfilled and which elements of the models contribute to it. Let us illustrate this point with Bueno's and Colyvan's work. They claim that "the fundamental role of applied mathematics is inferential" (Bueno and Colyvan, 2011, 352) and accordingly propose an "inferential conception" of the application of mathematics that extends Hughes' three-step DDI account of scientific representation (see below).[2] First, a "mapping from the empirical set up to a convenient mathematical structure" (ibidem, 353) is established (immersion step); by doing so, it becomes possible "to obtain inferences that would otherwise be extraordinarily hard (if not impossible) to obtain" (ibidem, 352) (derivation step); finally, the mathematical consequences that were obtained are interpreted step in terms of the initial empirical set up (ibidem, 353) (interpretation step). Bueno and Colyvan further highlight the importance of the inferential role of mathematics for mathematical unification, novel predictions by mathematical reasoning or mathematical explanations (ibidem, 363). However, the analysis of how this inferential role is carried out shines by its absence. Bueno and Colyvan mostly analyze mathematical resources in a semantic perspective[3] and insist on the difference in content and interpretation that these make possible, e.g., when "mathematics provides additional entities to quantify

[1] Frigg, while clearly stating the problem, does not really address it and is content with briefly emphasizing the advantages of his fictional account of model concerning the epistemology of models (Frigg, 2010). As to the epistemological section of Frigg and Hartmann's review article about scientific models, it merely points at experiments, simulations, thought-experiment as ways of investigating models (Frigg and Hartmann, 2017).

[2] Suarez's inferential conception (Suarez, 2004) hardly addresses either the question of how inferences from models are actually carried out. For lack of space, we shall not discuss it here.

[3] Their discussion is mostly directed at the shortcomings of Pincock's "mapping account" of the application of mathematics (Pincock, 2004).

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -92-

over" (complex numbers), or is "the source of interpretations that are physically meaningful" and provide "novel prediction" about physical systems, like with the case of the interpretation of negative energy solutions to Dirac's equation (ibidem, 366).

In another paper, Bueno suggests that results are derived "by exploring the mathematical resources of the model" in which features of the empirical set up are immersed (Bueno, 2014, 379, see also 387) and that results emerge "as a feature of the mathematics" (ibidem) or by using "the particular mathematical framework" (ibidem, 385). What this inferential power of mathematics should be specifically ascribed to remains unclear. Bueno and Colyvan (2011, 352) just claim that the "embedding *into a mathematical structure* makes it is possible to obtain inferences". They also emphasize how, with the help of appropriate idealizations, "the *mathematical model* [can] directly [yield] the results" (ibidem, 360, our emphasis). But elsewhere in the paper, consequences are said to be drawn "*from the mathematical formalism*, using the mathematical structure obtained in the immersion step" (ibidem, 353, our emphasis).

What are we to make of these various claims? A prima facie plausible answer to this question might be that structures and formalisms are the two sides of a same inferential coin. However, this answer is not satisfactory, since, as is well-known, mathematical structures can be presented in different formalisms, which, as we shall see in Section 4, are associated with different inferential possibilities. Another blind spot in Bueno's and Colyvan's account is that while the derivation step is claimed to be "the *key point* of the application process, where consequences from the mathematical *formalism* are generated'' (ibidem, 353), the question of how inferences are drawn with the help of formalisms is left under-discussed.

We draw from this brief analysis of Bueno's and Colyvan's views that the notions of mathematical resource and inferential power, which are commonly used when discussing applications of mathematics, are often mere labels in need of further investigation. Coming back to the seminal ideas presented by Hughes and extended by Bueno and Colyvan is of little help because Hughes' paper lacks precise answers to the following precise questions: What are exactly mathematical resources? What is their inferential power? In his DDI (Denotation, Demonstration, and Interpretation) account of scientific representation, Hughes claims that scientific representations have an ''internal dynamic'', whose effects we can examine (1997, 332), and "contain *resources* which enable us to demonstrate the results we are interested in". A general notion of resource is appropriate to capture the variety of ways in which demonstrations can be

Seattle, WA; 1-4 November 2018 -93-

carried out; however, the claim that the deductive power comes from "the *deductive resources* of mathematics they employ" (ibidem, 332) is too vague and is left unanalyzed.

3. Components of mathematical resources. How are the notions of inferential power and mathematical resources to be analyzed? Are they linked to structures or to symbolic systems and formalisms? In this section, we claim that formalisms are an important component of the notions of inferential power and mathematical resource and should be analyzed in their own right.

Let us begin by briefly presenting what are, according to us, the three main components of the notions of mathematical resource and associated inferential power. First, mathematical structures, *to the extent that they are tractable*, are undoubtedly an important part of the mathematical resources that are used in mathematical modeling. As argued by Cartwright, theories are no "vending machines" that "drop out the sought-for representation" (1999, 247); scientific models are no vending machines either and scientists must make the best of the models that they know to be tractable. Accordingly, the content of models often needs to be adapted by means of idealizations, approximations (Redhead 1980), abstractions, by squeezing representations into the straightjacket of a few elementary models (Cartwright, 1981), or by drawing, from the start, on the pool of existing tractable models (Humphreys, 2004, Barberousse and Imbert, 2014).

Second, mathematical knowledge associated with structures is also to be counted as a distinct mathematical resource, which allows for new inferences when it is available. Let us take the well-known example of Koenigsberg's seven bridges. The impossibility of crossing them once and only once in a single trip can be demonstrated by applying a result from graph theory. Similarly, the explanation of the life-cycle of the Magicicada (Baker 2009, Colyvan 2018) is provided by the application of a number-theoretic property of prime numbers to life-cycles of species.

At last, formal settings or formalisms provide languages in which theories are developed, calculations carried out, and inferences drawn from models. Examples of formalisms are Hamiltonian formalism, path integrals, Fourier representation, cellular automata, etc. We provide a detailed analysis of some of these below. Contrary to mathematical structures, formalisms are partly content neutral (though form and content are often intertwined in scientific representations). As providing a partially stan-

dardized way of making inferences, they are important tools for scientists, which in turn justifies considering them as important units of analysis in the philosophy of science. Other authors have started exploring the idea that format matters in scientific activities. Humphreys gives general arguments to this effect and emphasizes the difference between formats that are appropriate for human-made and format that suit computational inferences (2004). Vorms (2009) also emphasizes the general importance of formats of representation when toying with theories or models. Formalisms are a specifically mathematical type of format whose role needs further investigation. This is what we do in the next section.

4. What are formalisms? As briefly stated above, formalisms are mathematical languages that allow one to present mathematical statements or objects and draw inferences about them by means of general inference rules. For example, *Hamiltonian formalism* is one of the formalisms through which scientists may find out means to solve differential equations. *Path integrals* is another formalism of this kind, with the help of which one may also solve (partial) differential equations. Let us illustrate the latter point further: the integral solution of the Schrödinger equation requires using a mathematical object, the *propagator,* whose calculation the path integrals formalism makes easier. *Fourier representation or formalism* enables one to represent mathematical functions as the continuous sum of sine functions (or complex exponential functions), so that harmonic analysis, i.e. the decomposition of a signal in its harmonic frequencies, may be performed. It also provides modelers with a way to express the solutions of some partial differential equations, such as the heat equation. Finally, formalisms like *numerical integrators, cellular automata, lattice Boltzmann methods,* and *discrete variational integrators*, are indispensable in current computational science.

Formalisms consist in the following elements:

i. elementary symbols;

ii. syntax rules that determine the set of well-formed expressions;

iii. inference rules;

iv. a partly detachable interpretation, both mathematical and physical.

Their use is facilitated by

v. translation rules that indicate how to shift from one formalism to another.

Let us illustrate these elements by discussing in more detail the above examples. In the Hamiltonian formalism, elementary symbols are used for a variable and its conju-

gate momentum: “(q, p)”, or for Poisson brackets ‘’{.,.}”. Among the syntax rules that are specific to Hamiltonian formalism, some allow one to rewrite Hamilton equations by using the canonical variables. Inferences rules allow the users to use action-angle variables (I, *theta*) and to solve equations by using these coordinates because this change of variables opens the possibility to deal with integrable systems, thus providing a systematic method to solve *exactly*, i.e., in closed forms, differential systems like the simple pendulum, and more generally, any 1D-conservative system. Indeed, due to this change of variables, one takes full advantage of the existence of conserved quantities in mechanical systems, which are then used as variables (actions) in Hamilton equations. This allows constructing the solution of the equations by “quadrature” (Babelon et al. 2003, chapter 2). An example of a translation rule is the Legendre transform that allows one to shift to Lagrangian formalism. Similarly, in the case of Fourier transforms, an elementary specific symbol is $f^\wedge$, which corresponds to the Fourier transform of the function *f*. Scientists use sets of rules that describe the Fourier transforms of some typical functions, such as the constant function, the unit step function, and the sinusoids, but also rules for the convolution product, viz. the Fourier transform of the convolution $f \circ g$ is the product of Fourier transforms of *f* and *g*: $(f \circ g)^\wedge = f^\wedge . g^\wedge$, so that solutions of equations may be found within Fourier space. An inverse Fourier transform is also defined, which enables one to move back from the Fourier transform $f^\wedge$ to the function *f* (this is again a translation rule).

As emphasized above, formalisms are (partly) content neutral and thus “exportable”, even though they usually come with a privileged physical interpretation. As a matter of fact, most formalisms have been developed within a peculiar modeling context or are linked to a physical theory. From this origin, the most successful ones may become autonomous and depart from their original, physical interpretation. For example, Hamiltonian formalism was initially developed in the context of classical mechanics but is nowadays autonomous and used in other physical contexts. Path integrals originally come from the study of Brownian motion (Wiener 1923) and quantum mechanics (Feymann 1942) but are currently used in other fields like field theory and financial modeling.

The mathematical interpretation of formalisms may sometimes be detachable. For example, the transition rules associated with cellular automata (see below) do not have any obvious mathematical interpretation. Further, although some formalisms are linked to acknowledged mathematical theories (e.g., the Fourier formalism is linked to

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -96-

the theory of complex functions), they differ from genuine mathematical theories, as shown by the example of path integrals, in which the formalism is used in the absence of any uncontroversial mathematical theory that could back it up. The definition of a path integral:

$$K(b,a) = \int_a^b e^{\frac{2i\pi}{h}\int_{t_b}^{t_a} L\,dt}\, Dx(t)$$

requires using a measure "Dx", to which no general, rigorous definition can be given yet. This mathematical concern does not prevent physicists from using path integrals anyway, as testified by the following quote: "The question of how the path integral is to be understood in full generality remains open. Given this, one might expect to see the physicists expending great energy trying to clarify the precise mathematical meaning of the path integral. Curiously, we again find that this is not the case" (Davey 2003, 450).

Let us finally emphasize that formalisms also differ from formulations of physical theories and allow philosophers of science to address different philosophical problems. Formulations of theories, in particular axiomatic ones, are explored when questions about conceptual content and metaphysical implications are raised. They pertain to foundational issues. Whether a given formulation involves calculus is a peripheral issue in this context. By contrast, the primary virtue of a formalism is to allow modelers to draw actual inferences from a theory or model. The inferential rules it contains are more important than the mathematical rigor of the language in which it is expressed.

5. Choosing a formalism. So far, we have argued that the inferential power that is required to explore models is partly brought about by formalisms, and we have given examples thereof. Accordingly, formalisms have to be carefully examined by philosophers of science if they are to provide a fine-grained analysis of how scientific knowledge is produced in practice. We now aim to show that there is no unique description of formalism-rooted inferential power since different formalisms allow for different types of inferences and are adapted to different types of inquiries. We do so by providing examples of these differences and of the factors that guide scientists when choosing the formalism that is best suited to the task at hand.

How do scientists decide which formalism to use in a given inquiry? The choice may first depend on the type of models at hand. For example, the path integral formalism is

well adapted to solve systems with many degrees of freedom (Zinn-Justin 2009) and makes ''certain numerical calculations in quantum mechanics more tractable'' (Davey 2003, 449). Lagrangian formalism offers a well-suited framework to solve equations describing constrained systems (Goldstein 2002, 13, Vorms 2009, 15). Fourier representation allows one to solve, e.g., the differential equations describing the time evolution of electrical quantities in networks. In this case, differential equations are transformed into *algebraic equations* on variables in Fourier space, which may be easier to solve. Finally, with the change of action-angle variables, Hamiltonian formalism potentially provides exact solutions for integrable systems, which have as many independent conserved quantities as degrees of freedom.

The use of a particular formalism is also guided by epistemic goals. Depending on the chosen formalism, different kinds of properties, general (e.g. periodicity, symmetry) or particular (dynamical), may be inferred from the same model. Let us illustrate this point with the example of prey-predator models in ecology. Among these, some obey Lotka-Volterra (LV) equations and represent transforming populations with a system of two coupled equations. If they are investigated within the Hamilton formalism, *general properties* of these models can be found without setting initial conditions or numerical values for the involved parameters. The reframed models can indeed be shown to be integrable, like the simple pendulum in classical mechanics. Dutt explicitly emphasizes the advantages of using this formalism for a two-species LV system:

> ''In dealing with the problems involving *periodicity*, the Hamilton-Jacobi canonical theory has a distinct advantage over the conventional methods of classical mechanics. In this approach, one introduces action and angle variables through canonical transformations in such a way that the angle variable becomes cyclic. One then obtains the frequency of oscillation by taking the derivative of the Hamiltonian with respect to the action variable. One may thus *bypass the difficulty* in obtaining the complete solutions of the equations of motion, *if these are not required*.'' (Dutt, 1976, 460, our emphasis)

LV models can also be solved with the help of computers and generic numerical integrators when the aim is to obtain particular dynamics for specific values of parameters and initial conditions. Such numerical solutions of the LV model can also be provided by specific formalisms, such as discrete variational integrators (Krauss 2017, 34; Tyranowski 2014, 149). In that case, discrete equations are derived from a discrete least action principle, which is well-suited to conservative systems, like the LV sys-

tem. Discrete variational integrators allow for the preservation of general properties like the conservation of global quantities, viz. energy, momenta, and symplecticity. This discrete formalism comes with mathematical constraints on the discretization of time since the time step has to be adaptive in order to guarantee the conservation of global quantities (Marsden & West 2001, Section 4.1).

Finally, let us mention that LV models can also be studied by using *cellular automata* (CA) and associated formalism, with the following advantages:

> [a rather general predator-prey model] is formulated in terms of automata networks, which describe more correctly the *local character* of predation than differential equations. An automata network is a graph with a discrete variable at each vertex which evolves in discrete time steps according to a definite rule involving the values of neighboring vertex variables. (Ermentrout and Edemstein-Keshet 1993, 106)

On the one hand, CA are discrete dynamical systems, but on the other, they are also a nice means to practice science with the help of a computationally simple formalism (in terms of transition rules). They can be extremely powerful. For example, rule 110 is Turing complete and, like lambda-calculus, can emulate any Turing machine and therefore complete any computation. In contrast with the case of Hamilton formalism, CA-based inferences from prey-predator models are carried out for specific values and parameters. As CA are described by local rules, these inferences merely pertain to local variations in the model. However, the simplicity of these rules is a tremendous advantage for modeling and code-writing. For instance, CA allow one to easily add rules for the pursuit and evasion of populations as well as rules for age variation (Boccara et al. 1993, Ermentrout and Edemstein-Keshet 1993, see also Barberousse and Imbert 2013 for an analysis of CA as used in fluid dynamics and compared with Navier-Stokes based methods).

Let us now turn to a different example illustrating how different the epistemological effects of using this or that formalism may be. Crystals are currently modeled as lattices that come under two forms, *lattices in real space* and *lattices in reciprocal space*. Each is associated with a specific formalism. Within the *real space lattice* formalism, crystals are described with a vector R expanded on a vector basis (a_1, a_2, a_3) which corresponds to crystal directions, and *alpha*, *beta*, *gamma* are the corresponding angles. Inferences about *symmetry* of crystals are usually made within this type of representation since the real space is well adapted to studying discrete translations and rotations.

Crystals can also be described with the help of a vector R* in a *lattice in reciprocal space*. There is a clear correspondence between the two spaces since they are dual. Given R in the real space, we can derive R* in the reciprocal space, and conversely. The two spaces are related by a Fourier transform. However, the *reciprocal space* can be more convenient because inferences about *diffraction and interference patterns* are easier to carry out in the Fourier representation. As stressed by Hammond in a textbook of crystallography:

> the reciprocal lattice is the basis upon which the geometry of X-ray and electron diffraction patterns can be most easily *understood* and […] the electron diffraction patterns observed in the electron microscope, or the X-ray diffraction patterns recorded with a precession camera, are simply sections through the reciprocal lattice of a crystal (Hammond 2009, 165).

This example shows that facilitating inferences may have various epistemological effects. Some are relevant to computational aspects and the predictions or explanations that scientists are able to produce in practice. Others pertain to the way scientists understand and reason about models and their target systems. This example also shows how different epistemic goals (symmetry-oriented vs. interference-oriented investigations of crystals) determine which formalism is chosen.

Overall, the above shows that formalisms not only have an important impact on the amount of results scientists may produce, but also on the types of results that are attainable. The examples we have discussed also highlight that the existence of a variety of formalisms is a source of epistemic richness and enhanced inferential power for scientists because it provides them with multiple ways of investigating the same mathematical structures or structures that are related by suitable morphisms.

6. Conclusion. The above proposals are meant to contribute to the epistemological question of what provides models with inferential power and helps scientists succeeding in their inquiries. We have shown that some of this inferential power is brought about by the formal symbolic tools that scientists use to present and investigate mathematical models. Our second claim is that all formal settings do not enable the same types of inferences nor are suited to all epistemic goals. Accordingly, a fine-grained analysis of the conditions of scientific progress needs, among other things, to focus on formalisms.

Our epistemological analysis is not tied to any particular theory of scientific representation. However, by showing that inferences actually hinge on choice of formalism, it suggests that a theory of scientific representation that is cashed out in terms of structures is too abstract to account for the various ways equations are solved in practice and information extracted from scientific models.

References

Babelon Olivier, Bernard Denis, and Talon Michel. 2003. *Introduction to classical integrable systems*, Cambridge: Cambridge University Press.

Baker, A. 2009. "Mathematical Explanation in Science". *British Journal for the Philosophy of Science* 60 (3): 611–633.

Barberousse, Anouk, and Cyrille Imbert. "New Mathematics for Old Physics: The Case of Lattice Fluids." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 44 (3) : 231–41.

Barberousse, Anouk, and Cyrille Imbert. 2014. "Recurring Models and Sensitivity to Computational Constraints" *The Monist* 97 (3): 259–79.

Boccara Nino, Roblin O. and Roger Morgan. 1994. Automata network predator-prey model with pursuit and evasion, *Physical Review E* 50 (6): 4531–41

Bueno, Otávio, and Mark Colyvan. 2011. "An Inferential Conception of the Application of Mathematics". *Noûs* 45 (2): 345–74.

Bueno, Otávio. 2014. "Computer Simulations: An Inferential Conception". *The Monist* 97 (3): 378–98.

Cartwright, Nancy (1983). *How the Laws of Physics Lie*. Clarendon Press, Oxford.

Cartwright, Nancy.1999. "Models and the Limits of Theory: Quantum Hamiltonians and the BCS Models of Superconductivity". In *Models as Mediators*, ed. Mary S. Morgan and Margaret Morrison Morgan, Cambridge: CU Press: 241–81.

Colyvan, Mark. Forthcoming. "The Ins and Outs of Mathematical Explanation", *Mathematical Intelligencer*.

Davey Kevin. 2003. "Is Mathematical Rigor Necessary in Physics?" *The British Society for the Philosophy of Science*, 54(3): 439–463

Dutt Ranabir. 1976. "Application of the Hamiltonian-Jacobi Theory to Lotka-Volterra Oscillator", *Bulletin of Mathematical Biology*, 38: 459–465.

Ermentrout G. Bard and Edemstein-Keshet, Leah. 1993. "Cellular Automata Approaches to Biological Modeling". *Journal of Theoretical Biology* 160: 97–133.

Feynman, Richard. P. 1942. "The Principle of least action in quantum mechanics", *PhD. diss.*, Princeton University.

Frigg, Roman. 2010. "Models and Fiction". *Synthese* 172 (2): 251–68.

Frigg, Roman, and Stephan Hartmann. 2017. "Models in Science." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/models-science/.

Goldstein, Herbert. 2002. *Classical Mechanics*. Reading, Mass: Addison-Wesley.

Hammond, Christopher. 2009. *The Basics of Crystallography and Diffraction*, Oxford University Press.

Hughes, Robert I.G. 1997. "Models and Representation". *Philosophy of Science* (Proceedings): 64: S325–S336.

Hughes, Robert I.G. 2010. *The Theoretical Practices of Physics: Philosophical Essays*. Oxford: Oxford University Press.

Humphreys, Paul. 2004. *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Kraus, Michael. 2017. "Projected Variational Integrators for Degenerate Lagrangian Systems", preprint: https://arxiv.org/pdf/1708.07356.pdf

Marsden Jerrold E. and West Matthew. 2001. "Discrete Mechanics and Variational Integrators", *Acta Numerica*, 10: 357–514.

Morgan, M., and Margaret Morrison (1999). *Models as Mediators*. Cambridge University Press.

Pincock, Christopher. 2004. "A new perspective on the problem of applying mathematics", *Philosophia Mathematica* 3 (12), 135-61.

Redhead, M. 1980. "Models in Physics", *The British Journal for the Philosophy of Science*, 31(2): 145-163

Suarez, Mauricio. 2002. "An Inferential Conception of Scientific Representation", *Philosophy of Science* 71 (5): 767-779

Tyranowski Tomasz. M. 2014. "Geometric integration applied to moving mesh methods and degenerate Lagrangians". Ph.D. diss., California Institute of Technology.

Vorms, Marion. 2011. "Formats of Representation in Scientific Theorizing." In *Models, Simulations, and Representations*, edited Paul Humphreys and Cyrille Imbert. Routledge.

Wiener, Norbert. 1923. "Differential space". *Journal of Mathematical Physics* 2: 131-174.

Zinn-Justin Jean. (2009), Path Integral, *Scholarpedia*, 4(2): 8674.

Representation Re-construed: Answering the Job Description Challenge with a Construal-based Notion of Natural Representation

Abstract: Many philosophers worry that cognitive scientists apply the concept REPRESENTATION too liberally. For example, William Ramsey argues that scientists often ascribe natural representations according to the "receptor notion," a causal account with absurd consequences. I rehabilitate the receptor notion by augmenting it with a background condition: that natural representations are ascribed only to systems construed as organisms. This Organism-Receptor account rationalizes our existing conceptual practice, including the fact that scientists in fact reject Ramsey's absurd consequences. The Organism-Receptor account raises some worrying questions, but as a more faithful characterization of scientific practice it is a better guide to conceptual reform.

Abstract: 100 words

Total: 4,995 words

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -104-

1. Introduction. There is a common complaint among philosophers that scientists use the word "representation" too liberally. Representation is often contrasted with indication: representation is a distinction achieved by maps, linguistic performances, and thoughts, whereas indication is a less-demanding state achieved by thermostats, which indicate ambient temperature, and refrigerator lights, which indicate whether the door is open (Dretske 1981; Cummins and Poirier 2004). However, cognitive scientists often ascribe representations when it seems that mere indication is all that is called for. We commonly say that hidden layers in a neural network represent concepts, or that neurons in V1 represent visual edges, because they reliably respond differently to the circumstances they are said to represent (Ramsey 2007, 119–20; cf. Hubel and Wiesel 1962). But these "representations" are thin-blooded compared to paradigmatic conventional representations. For example, they cannot be invoked in the absence of an appropriate stimulus. So are cognitive scientists conceptually confused? Do they exaggerate their claims? And if the natural representations posited by cognitive scientists aren't genuine representations, is the cognitive revolution dead?

William Ramsey provides an excellent book-length exploration of these worries, articulating a qualified pessimism about their answers:

> …we have accounts that are characterized as "representational," but where the structures and states called representations are actually doing something else. This has led to some important misconceptions about the status of representationalism, the nature of cognitive science and the direction in which it is headed. (2007, 3)

Ramsey describes the "job description challenge": to give an account of the distinctive properties of representations in virtue of which appealing to them serves a special

explanatory role. If the job description challenge can be met, then we can formulate a plan for conceptual reform.

I undertake Ramsey's challenge, but with a metadiscursive twist: I describe the Organism-Receptor account, which articulates conditions for ascribing representations, in virtue of which such ascriptions achieve a special explanatory purpose. The account is merely suggestive about the properties that distinguish first-order representational states from non-representational states; it says more about the mental state of the ascriber than about the representation-bearing system. However, the Organism-Receptor account provides a more adequate characterization of scientists' practice than Ramsey's.

My main aim in this paper is to push back against pessimistic evaluations of the existing practice of representation-ascription in cognitive science, like Ramsey's. I will focus on Ramsey's critique of the "receptor notion," a flawed causal theory of representation that he attributes to some cognitive scientists. Ramsey argues that the receptor notion has absurd consequences, although scientists do not accept them. By augmenting the receptor notion with a construal-based background condition, I can explain why scientists do not draw these absurd conclusions. Whereas Ramsey's pessimistic account of scientists' practice of ascribing representations finds it wanting and is extensionally inadequate, mine rationalizes our extant conceptual practice (though that practice is not beyond criticism). I conclude that my apologetic account is a more charitable and adequate interpretation of existing scientific practice than Ramsey's.

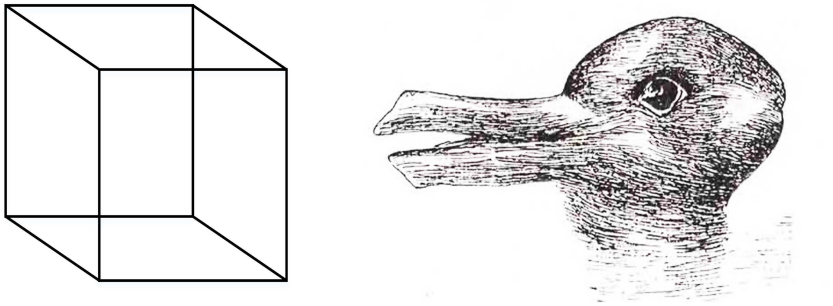
2. Ramsey on the "Receptor Notion." Ramsey argues that natural representations in cognitive science are often ascribed according to the "receptor notion," a crude causal theory of representation. According to the receptor notion, a state *s* represents a state of affairs *p* if *s* is regularly and reliably caused by *p* (2007, 119).

Ramsey claims that the receptor notion is what justifies the ascription of representations to cells in V1 that detect visual edges, cells in frog cortex that detect flies, and the mechanisms in Venus flytraps that cause their "jaws" to close (119–23). Ramsey argues that this receptor notion is too liberal to be useful to scientists. For example, it is susceptible to the "disjunction problem" (Fodor 1987): since frog neurons respond reliably to visual stimulation by flies *or* (say) BBs, we should say that the content of the representation is *fly-or-BB*, rather than *fly*. Likewise, Venus flytraps represent objects in a particular range of sizes rather than *edible insects*, and the human concept GOAT represents *goats-or-weird-looking-sheep*. Such disjunctive content-ascriptions are usually considered absurd. Absent a clever fix, we must embrace unwieldy, disjunctive contents for representations or we must reject the receptor notion (Ramsey, 129).

Dretske's (1988) teleofunctional theory of representation is a sophisticated twist on the receptor notion that avoids the disjunction problem. On Dretske's view, a representational state must not only be causally dependent on the state of affairs it represents, but must serve a function for its containing system in virtue of this causal dependency. This extra condition motivates constraints on representational content that eliminate problematic disjunctive contents. Dretske's theory is subject to some subtle criticisms that I will discuss in Section 6, but the Organism-Receptor account will preserve some of the teleological character of Dretske's theory.

Ramsey's most compelling objection to the receptor account, including Dretske's sophisticated version, is that it justifies ascribing representational contents to states that are not, in fact, representational: smoke "represents" fire since the latter causes the former. Likewise, the firing pin of a gun "represents" whether the trigger is depressed, and rusting iron "represents" the presence of water and oxygen (138–47). Ramsey claims, plausibly, that these are absurd consequences. I find Ramsey's reductio

Seattle, WA; 1-4 November 2018 -107-



Ambiguous figures. Left: The Necker cube. Right: The duck-rabbit (image from Jastrow 1899).

compelling, but reject a different premise than he does. Rather than conclude that cognitive scientists have a bad conceptual practice, I question whether his characterization of the receptor notion is a charitable understanding of what happens in cognitive science. After all, cognitive scientists do not generally claim that GOAT denotes *goats-or-sheep* (at least for competent judges of goathood), or that firing pins represent anything.

3. A Construal-based Notion of an Organism. I argue that something like the receptor notion can be salvaged if being a receptor is contextualized in terms of construal. Construal (also called "seeing-as") is a judgment-like attitude whose semantic value can vary licitly independently of the state of affairs it describes. For example, we can construe an ambiguous figure like the Necker cube as if it were viewed from above or below, or the duck-rabbit as if it were an image of a duck or of a rabbit (Roberts 1988; see also Wittgenstein 1953). We can construe an action like

skydiving as brave or foolhardy, depending on which features of skydiving we attend to.

On a construal-based account of conceptual norms, a concept (e.g. REPRESENTATION) is ascribed relative to a construal of a situation. For example, perhaps I fear something only if I construe it as dangerous to me or detrimental to my ends (Roberts 1988). Daniel Dennett's (1987) intentional stance is a more familiar example: according Dennett, a system has mental states if and only if we construe it in such a way that its behavior is explainable in terms of a belief-desire schema.

I propose that construing something as an organism involves construing it such that it has goals and behavior, and believing that it has mechanisms that promote those goals by producing that behavior. More precisely:

> **Organism-Construal.** A subject *a* construes a system *x* as an organism in a context[1] *c* if and only if, in *c*,
>
> (O1) *a* attributes a set of goals *G* to *x*,
>
> (O2) *a* attributes a set of behaviors *B* to x,
>
> (O3) *a* believes that the elements of *B* function to promote elements of *G*,
>
> (O4) *a* believes that *x* possesses a set of mechanisms *M*, and
>
> (O5) *a* believes that the elements of *M* collectively produce the elements of *B*.

My main argument does not rely on all the details of Organism-Construal; it could be replaced by a different explication of what it is to see something as an organism. But Organism-Construal captures an intuitive notion of a critter. First of all, we normally take living critters to have goals, such as survival and reproduction, and behaviors that

[1] The relevant notion of a context is something like MacFarlane's (2014) "context of assessment."

promote those goals. However, Organism-Construal does not require that an organism really have goals (whatever that involves) or exhibit behavior (however that's distinguished from other performances). To see something as an organism according to Organism-Construal, the construing subject need only *attribute* goals to the system, and see some of its performances as behaviors that promote those goals. Such goals could include relatively specific aims such as locating food, getting out of the rain, or driving home. We sometimes also attribute goals and behaviors to non-living things, such as automated machines. For example, we might say that a robot vacuum has the goal of cleaning the floor, which it accomplishes by sucking up dust. Or I might say that my GPS navigation computer is trying to kill me, which it accomplishes by consistently giving me directions that lead me through strange, dangerous backroads. Condition (O3) is expressed in terms of belief instead of attribution, meaning that the construing subject must sincerely believe that an organism's putative behaviors function to promote its putative goals. When and insofar as someone construes a system in this way, the conditions (O1)–(O3) above are satisfied.

Conditions (O4)–(O5) require that the system's behavior be explainable by appeal to mechanisms. "Mechanisms" here should be understood in roughly the sense meant by the new mechanists (Machamer, Darden, and Craver 2000; Bechtel and Abrahamsen 2005; Craver 2007): organized structures of component parts and operations that produce a phenomenon, and the description of which is an explanatory aim of some scientific projects. Much explanation in biology and neuroscience plausibly follows a mechanistic model, and likewise in cognitive science. Daniel Weiskopf (2011) has argued that cognitive explanations are not properly mechanistic, but even on his view cognitive explanations are extremely similar to mechanistic ones, distinguishable only because the relationship between components of cognitive models and their physiological realizers is relatively opaque. Regardless, cognitive scientists use the word "mechanism" to refer to the referents of their models,

just as biologists and neuroscientists do. I am more moved by the similarities between the biological and the cognitive sciences than the differences. Therefore, like Catherine Stinson (2016), I acknowledge Weiskopf's concerns but nevertheless adopt the language of "mechanisms."

Not all of a system's mechanisms function to produce behavior. For example, biological organisms have metabolic and other mechanisms that maintain bodily integrity. Such mechanisms may need to function correctly as a background condition for the organism to behave, but scientists do not typically take behavioral patterns to be the explanandum phenomena of such mechanisms. Let us call mechanisms that do contribute to the explanation of behavior *behavioral mechanisms*. As for what it means for a system to "possess" a mechanism, a mereological criterion will do for now: the mechanism must be a part of the system. Condition (O5) is meant to limit the mechanisms in the set *M* to behavioral mechanisms.

So far so abstract; let's consider an example. The robot Herbert was designed to wander autonomously through the MIT robotics lab, avoiding obstacles, and collecting soda cans with its arm (Brooks, Connell, and Ning 1988). Herbert can be construed as an organism, even though it is not alive, as long as one (O1) attributes goals, like avoiding collisions and collecting soda cans, to Herbert, (O2) sees some of Herbert's performances as behaviors, (O3) believes that Herbert's behaviors promote its goals, and (O4) believes that Herbert possesses mechanisms that (O5) explain its behavior. Herbert does possess mechanisms for accomplishing goals; it is equipped with sensors, computers, and motors that coordinate its locomotion and its grasping arm. And most people readily anthropomorphize Herbert enough to see it as a goal-directed, behaving system (pace Adams and Garrison [2013], who insist that Herbert has its designers' goals, but no goals of its own). Anyone willing to engage in the imaginative attribution of goals and behavior to Herbert can see Herbert as an organism, even if on reflection they believe Herbert is not literally an organism. The

willingness to ascribe representations to a system plausibly waxes and wanes along with one's willingness to construe the system as an organism in something like the sense described above. There are psychological limits on the willingness to attribute goals and behaviors to systems relatively unlike animals, and these limits may vary between individuals.

4. The Receptor Notion Re-construed. Returning now to the receptor notion of natural representation, I suggest that it can be augmented in the following way:

> **Organism-Receptor.** A state *s* represents a state of affairs *p* if
> (R1) *s* is regularly and reliably caused by *p*, and
> (R2) *s* is a functional state of a behavioral mechanism possessed by an organism.

Organism-Receptor is not a construal-based explication, but it depends on a construal-based account of ORGANISM. It preserves the spirit of Ramsey's receptor notion, with the added condition that representations be ascribed to parts of systems construed as organisms. Representation-ascriptions guided by Organism-Receptor inherit their plausibility from the plausibility of the corresponding construal of some system as an organism. Most accounts of cognitive representation require there to be a representational subject of some kind (e.g. Adams and Aizawa 2001; Rupert 2009; Rowlands 2010), and on Organism-Receptor the organism serves this role. We can constrain the acceptable contents of these representations by requiring they correspond to descriptions of *p* according to which *p* is relevant to the pursuit of an organism's goals. This appeal to goals is not ad hoc, since according to Organism-Receptor representations are ascribed to organisms, i.e. systems to which we've already attributed a set of goals. Thus, like Dretske's (1988) and Millikan's (1984)

teleofunctional accounts, this construal-based account addresses the disjunction problem by appealing to goals of organisms.

The metadiscursive job-description challenge is to provide criteria of ascription for representations, in virtue of which representation-ascriptions achieve some explanatory purpose. I have provided criteria of ascription, so what is their purpose? On Donald Davidson's (1963, 5) account of intentional action, actions are performed under the guise of a privileged description (or set of descriptions). Davidson flips the light switch in order to turn on the light, but not in order to alert the prowler outside (whose presence is unknown to Davidson) that he is home, though he also does the latter. Davidson calls this feature of action its "quasi-intensional character." Behavioral mechanisms also have something like a quasi-intensional character, since there are privileged descriptions that make explicit how they and their components contribute to an organism's capacity to pursue its goals. For example, edge-detecting cells in V1 fire in order to identify boundaries in an organism's environment, not to consume glucose, though they also do the latter. The use of representation-talk by cognitive scientists, as licensed by Organism-Receptor, is a way to habitually mark these privileged descriptions and distinguish them from other descriptions of the same states or events. And since cognitive science is concerned with the functional structure of behavior-coordinating mechanisms rather than other features of cognitive systems, it is easy to see why representation—even in this relatively thin sense—has always been the dominant theoretical perspective in cognitive science. This focus on quasi-intensional characterization may even be what makes the cognitive scientific perspective distinctive (on scientific perspectives, see e.g. Giere 2006).

The Organism-Receptor account provides us with resources to salvage the receptor notion from Ramsey's reductio. It is plausible to suppose that cognitive scientists generally ascribe natural representations to systems against an imaginative

Seattle, WA; 1-4 November 2018 -113-

background like this. After all, most cognitive science concerns the mechanisms of living systems, especially animals (except in computer science and some computational modeling, where the object of attention is a formal object like a connectionist network that is presumed to be analogous in some way to such a mechanism). Such systems are easily construed as organisms in the sense of Organism-Construal. Non-living things and even non-animals are in general more difficult to construe as organisms in that sense, since they are often perceived to lack goals, the capacity to behave, or both.

5. The Organism-Receptor Notion in Context. Consider a strong case of representation, like fly-detecting cells in frog visual cortex. We construe frogs as systems that exhibit goal-directed behavior and believe they possess mechanisms that explain that behavior. Frog visual cortex contains mechanisms that (along with other mechanisms) explain behaviors like fly-catching. When we identify cells in frog visual cortex that fire in response to the visual presence of flies (or fly-like objects), we ascribe representational properties to those cells. The contents we ascribe to representations in frog visual cortex are constrained by the goals we attribute to frogs. *That a small insect is present* is a suitable content because flies can be consumed for energy; *that a wiggly BB is present* does not have this significance for frogs, although BBs may be indistinguishable from insects by the mechanisms in the frog's visual cortex. Nevertheless, the relationship between fly-presence and the frog's goals provide a ground for privileging non-disjunctive descriptions of representational content.

The Organism-Receptor account also explains why liminal cases of representation, like the case of Herbert, are liminal. We can say that Herbert represents such states of affairs as the presence of obstacles and soda cans, because states of Herbert's sensors are regularly and reliably caused by those states of affairs.

And we can ascribe contents to representations by drawing on descriptions of Herbert's environment that relate to the goals we ascribe to Herbert. However, our willingness to take these representations seriously as natural representations that bear content intrinsically covaries with our willingness to take Herbert seriously as an organism. We are not as comfortable attributing genuine goals and behaviors to Herbert as we are attributing goals and behaviors to frogs.[2]

Finally, absurd cases like the firing pin can be excluded (for the most part) since guns are not easily construed as "organisms." Firearms are difficult to anthropomorphize, since they do not exhibit autonomous behavioral dynamics and we don't normally see them as having goals of their own. It is not *impossible* to ascribe goals to weapons or other tools, but the ascription of folk-psychological properties to tools, like the folk ascription of a bloodthirsty disposition to a sword, generally depends on the way a tool influences its users' behavior. (I suspect this dependence might offer some novel explanations of why Clark and Chalmers' [1998] extended cognition hypothesis is attractive to some.) The attribution of autonomous behaviors to tools like swords is fanciful. Perhaps we might imagine a tool exhibits psychic "behavior," but anyway we do not believe that swords possess mechanisms that produce this "behavior" (though if we did, such a construal would be more compelling). If the firing pin of a gun is not a component of a behavioral mechanism, it cannot represent anything according to the Organism-Receptor account.

So the Organism-Receptor account licenses an ascriptive practice that resembles the crude receptor notion when the role of construals is not made explicit. It is unusual in that it inverts Ramsey's preferred order of ascription: Ramsey wishes to

---

[2] Notably, Rodney Brooks himself does not claim that it is proper to ascribe representational capacities to Herbert (Brooks, Connell, and Ning 1988; Brooks 1991), but Brooks plausibly had in mind a more demanding account of representation.

Seattle, WA; 1-4 November 2018 -115-

ascribe cognitive structure to systems in virtue of their representational structure (see e.g. Ramsey, 222–235), whereas I suggest that we in fact ascribe representational structure in virtue of seeing a system as a system with goal-directed behavior, i.e. as a potentially cognitive system.

6. Worries. Since the Organism-Receptor account shares a certain teleological character with Dretske's account, I will discuss Ramsey's two most developed objections to Dretske, along with other worries specific to the Organism-Receptor account. First, Ramsey objects that Dretske's account is question-begging with regard to the job-description challenge. Roughly, teleological normativity (i.e. functioning and malfunctioning) is not sufficient to explain intentional normativity (i.e. representation and misrepresentation), and since Dretske provides no satisfying criteria for what it is for a state to function as a representation, he cannot bridge that gap (Ramsey 2007, 131–2). But the Organism-Receptor account has more resources than Dretske's teleofunctionalism. Construing a system as an organism involves construing it as exhibiting behavior, which allows us to distinguish behavioral mechanisms from other mechanisms. On the Organism-Receptor account, misrepresentations are malfunctions of behavioral mechanisms (like frog vision), but not of other mechanisms (like a frog's circulatory system or a gun's firing mechanism).

My reply invites a rejoinder: on the Organism-Receptor account the functional roles of representations will be extremely diverse, and representations will be common. They will not just include IO-representation and S-representation (roughly, information-processing relata and models for surrogative reasoning; Ramsey 2007, 68ff.), which Ramsey and most cognitive scientists regard as genuinely representational. They will also include more controversial varieties of "representation," such as Millikan's (1995) "pushmi-pullyu" representations: Janus-faced mechanistic components that simultaneously indicate a state of affairs and cause

an adaptive or designed response. In other words, representations will include what Ramsey calls "causal relays" like the firing pin in a gun, the inclusion of which in the extension of REPRESENTATION was the ground for his reductio! However, the absurd cases can be avoided. The firing pin case is excluded because guns are poor examples of organisms. And pushmi-pullyu representations include cases with significant intuitive appeal to many scientists, like the predator calls of vervet monkeys (Millikan 1995; cf. Seyfarth, Cheney, and Marler 1980). While this conception of representation has a more liberal extension than Ramsey is comfortable with, it is liberal enough to explain common representation-ascriptions in cognitive science without being so liberal as to countenance absurd cases like Ramsey's firing pin, so I submit it is adequate to scientific practice.

Ramsey's second objection is that Dretske is committed to a false principle: that if a component is incorporated into a mechanism because it carries information, then its function is to carry information (132–9). However, the Organism-Receptor account constrains the causal dependence criterion (R1) by relying on construals of systems as organisms instead of teleofunctional commitments. The account I describe is not committed to Dretske's principle, and therefore is not subject to this objection.[3]

Nevertheless, one might worry whether the organism criterion (R2) is a suitable condition on representation-ascription. I suggested five conditions (O1)–(O5) on what can be seen as an organism, but conditions (O1) and (O2) are fairly unconstrained. There are psychological limitations on when goals or behaviors can be plausibly attributed to a system, but what are those limits? And what factors influence interpersonal variability in willingness to make these attributions? The reason this practice isn't bonkers is that it coheres with the explanatory purpose of

[3] Ramsey's discussion is rich and worthy of deeper engagement than this, but for reasons of space I leave the matter here.

Seattle, WA; 1-4 November 2018 -117-

representation-ascriptions: to make explicit the quasi-intentional character of behavioral mechanisms. Nevertheless, we should hope that these psychological limitations are vindicated by more principled considerations. Criticism is warranted if scientists attribute goals and behaviors when they should not. There is some extant work on the proper norms ascribing goals to organisms (e.g. Shea 2013; Piccinini 2015, chap. 6), but little serious work on how to understand the concept of BEHAVIOR in the context of cognitive science. We should worry about the practice of ascribing natural representations if scientists construe things that are not cognitive systems as "organisms." Indeed, we might indeed worry that many cognitive scientists misuse the concept COGNITION, given the intense disagreements over its extension (see e.g. Akagi 2017). However, my present aim is not to evaluate scientific practice, but to describe it faithfully (with the hope that a more satisfactory evaluation will follow).

Another worry about construal-based accounts is that they entail an unattractive anti-realism: if representations and their contents only exist relative to construals, they are mind-dependent rather than objective, right? This worry is unfounded. I am undertaking a modified version of Ramsey's job description challenge: my aim is to describe the ascription of representations in virtue of which they serve an explanatory purpose, not to distinguish genuinely representational states from non-representational states. The Organism-Receptor account does not entail that representations exist relative to construals, only that they are *ascribed* relative to construals. My account is consistent with the existence of a first-order account of the metaphysics of representation that justifies this practice (or doesn't). After all, the duck-rabbit can be construed as a duck even if it is not a duck, and nothing about that fact entails that ducks (or unambiguous images of ducks) are not real. The Organism-Receptor account describes a norm that plausibly guides human scientists with imperfect capacities for knowledge. But while my solution to the metadiscursive job description challenge is not inconsistent with Ramsey's solution to

the first-order job description challenge, it is inconsistent with Ramsey's characterization of scientific norms for ascribing natural representations.

7. Conclusion. I began by observing the common worry that scientists ascribe representations more liberally than many philosophers are comfortable with, and in particular that scientists rely on an unsatisfactory "receptor" criterion. I sketched an account on which scientists ascribe natural representations only to components of mechanisms of systems construed as "organisms." Since in practice cognitive scientists attend almost exclusively to systems that are easily so construed, their behavior may appear to be guided by the crude receptor criterion whereas in fact it is guided by the Organism-Receptor criterion. However, while the Organism-Receptor account is still relatively liberal, a crucial difference between the two accounts is that the crude criterion has absurd consequences, whereas such consequences are eliminated or marginalized on the Organism-Receptor criterion. Since scientists do not in fact endorse these absurd consequences, I argue that the augmented criterion is a better hypothesis regarding norms for representation-ascription in cognitive science.

This is proposal is not a comprehensive, new theory of representation, but it accomplishes two things. First, it provides argumentative resources for resisting the common worry that cognitive scientists use hopelessly liberal criteria for ascribing representations. Second, it offers a novel picture of practices for representation-ascription in the biological and behavioral sciences, one that is less pessimistic picture than Ramsey regarding conceptual rigor in cognitive science. The picture is not beyond criticism—in particular, it wants for a more detailed account of the grounds that warrant attributing behaviors and goals to systems. But since it is more faithful to our practice than Ramsey's it is likely to yield more productive suggestions for how to guide that practice into the future. I suggest that we safeguard conceptual rigor in cognitive science not by cleaving more faithfully to the representationalism of the

cognitive revolution, but by embracing role of construal in scientific inquiry, making it explicit, and subjecting it to reasoned criticism.

REFERENCES

Adams, Fred, and Ken Aizawa. 2001. "The Bounds of Cognition." *Philosophical Psychology* 14:43–64.

Adams, Fred, and Rebecca Garrison. 2013. "The Mark of the Cognitive." *Minds and Machines* 23:339–52.

Akagi, Mikio. 2017. "Rethinking the Problem of Cognition." *Synthese*. doi: 10.1007/s11229-017-1383-2.

Bechtel, William, and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421–41.

Brooks, Rodney. 1991. "Intelligence without Representation." *Artificial Intelligence* 47:139–59.

Brooks, Rodney, Jonathan Connell, and Peter Ning. 1988. "Herbert: A Second Generation Mobile Robot." *A.I. Memos* 1016:0–10.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58:7–19.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Cummins, Robert, and Pierre Poirier. 2004. "Representation and Indication." In *Representation in Mind: New Approaches to Mental Representation*, E sds. Hugh Clapin, Phillip Staines and Peter Slezak, 21–40. Amsterdam: Elsevier.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* 60:685–700.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

———. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT/Bradford.

Giere, Ronald N. 2006. *Scientific Perspectivism*. Chicago: University of Chicago Press.

Hubel, David H., and Torsten N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160:106–54.

Jastrow, Joseph. 1899. "The Mind's Eye." *Popular Science Monthly* 54:299–312.

MacFarlane, John. 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Clarendon.

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67:1–25.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

———. 1995. "Pushmi-Pullyu Representations." *Philosophical Perspectives* 9:185–200.

Piccinini, Gualtiero. 2015. *Physical Computation: A Mechanist Account*. Oxford: Oxford University Press.

Ramsey, William M. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.

Roberts, Robert C. 1988. "What Emotion Is: A Sketch." *Philosophical Review* 97:183–209.

Rowlands, Mark. 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: MIT Press.

Rupert, Robert. 2009. *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.

Seyfarth, Robert M., Dorothy L. Cheney, and Peter Marler. 1980. "Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication." *Science* 210:801–3.

Shea, Nicholas. 2013. "Naturalising Representational Content." *Philosophy Compass* 8:496–509.

Stinson, Catherine. 2016. "Mechanisms in Psychology: Ripping Nature at Its Seams." *Synthese* 193:1585–614.

Weiskopf, Daniel A. 2011. "Models and Mechanisms in Psychological Explanation." *Synthese* 183:313–38.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. 3rd Ed. Trans. G.E.M. Anscombe. Eds. G.E.M. Anscombe and Rush Rhees. Oxford: Blackwell, 2001.

Comparing Systems Without Single Language Privileging

Max Bialek

mbialek@rutgers.edu

For the 2018 PSA Meeting.

Word count: 4753

Abstract

It is a standard feature of the BSA and its variants that systematizations of the world competing to be the best must be expressed in the same language. This paper argues that such single language privileging is problematic because (1) it enhances the objection that the BSA is insufficiently objective, and (2) it breaks the parallel between the BSA and scientific practice by not letting laws and basic kinds be identified/discovered together. A solution to these problems and the ones that prompt single language privileging is proposed in the form of privileging the best system competition(s).

1 Introduction

According to the Best Systems Analysis (BSA), the laws of nature are the theorems of the best systematization of the world—with 'best' standardly understood to mean the simplest and most informative (on balance). It is currently a standard feature of the BSA (since Lewis 1983) and its variants (Loewer 2007; Schrenk 2008; Cohen and Callender 2009) that a single language must be privileged as the language in which all systems competing to be the best will be expressed. Two problems have led these authors to adopt single language privileging: The first is the Trivial Systems Problem (TSP), according to which, in brief, allowing for suitably gerrymandered languages can guarantee that the "best" system will have axioms and theorems undeserving of the name "law" (see Lewis 1983 for its initial development). Language privileging provides a quick fix to the TSP as long as the privileged language is not among the suitably (and problematically) gerrymandered. The second is the Problem of Immanent Comparisons (PIC) suggested by Cohen and Callender (2009). The PIC takes it to be the case that there are only "immanent" measures for simplicity, strength, and their balance—that is, measures defined for only one language. With single language privileging, no two systems ever need to be compared when expressed in different languages, and so having to use only immanent measures is not an issue.

Though single language privileging solves these problems for the BSA and its variants, it creates new ones of its own. For one, the BSA is already often criticized for being insufficiently objective—because it is unclear that there is an objective answer to the question of what makes a system the best—and single language privileging has the potential to fuel those criticisms by requiring proponents of the BSA to say which

language gets privileged. Relativizing laws to languages (as in Schrenk 2008 and Cohen and Callender 2009) goes some way to resist such criticisms, but, as Bialek (2017) argues, relativity itself should be minimized (as much as scientific practice allows) when responding to those who employ the 'insufficiently objective' critique of the BSA. Another issue with language privileging—a version of which is suggested in a specific critique of Lewis (1983) by van Fraassen (1989), and is here newly generalized as an issue for *any* single language privinleging—is that it breaks the supposedly close connection in scientific practice between the discovery of the laws and the discovery of basic kinds.[1]

Both problems are, ultimately, overstated, and may be resolved not with single language privileging, but with the privileging of *classes* of languages. This addresses both of the issues just raised. For one, it restores the co-discovery of laws and basic kinds to the BSA by making the search for laws (via a best system competition conducted in the course of scientific practice) include a search through a class of languages for the one that yields the best system-language pair. It also helps to limit the degree to which laws may need to be relativized to language by reducing the problem of privileging a language (class) to the already present problem of choosing a measure of 'best'.

The outline of this paper is as follows. I begin, in Section 2, by laying out the PIC. In Section 3, I argue that the PIC ignores the existence of measures (illustrated by the

[1]Depending on the specific interests of the author, there has been talk of "basic kinds" (as in Cohen and Callender 2009), "fundamental kinds" (Loewer 2007), and "perfectly natural predicates" (Lewis 1983). These are progressively more restrictive ways of interpreting the predicates of a language that appear in the axioms of a best system expressed in that language. Throughout the paper I use the more general phrase "basic kinds", but nothing about that usage precludes a more restrictive reading.

Akaike Information Criterion) that, while not transcendent (since they cannot compare systems expressed in *any* two languages), are also not immanent (since they can compare systems expressed in *some* different languages). Being sensitive to the existence of such measures suggests a slightly different problem of *transcendent* measures, which may be resolved through privileging classes of languages. The problem for single language privileging of breaking the connection between the discovering laws and basic kinds is developed in Section 4, and its resolution via language-class privileging is demonstrated. In Section 5, I argue that the question of which language class to privilege is reducible to the question of which measure(s) of 'best' (simplicity, informativeness, etc.) should be used. Lastly, in Section 6, I note that the reducibility just introduced suggests a new solution to the TSP that is focused on choosing appropriate measures of 'best', with the conclusion being that none of the problems that have prompted language privileging actually require it for their resolution.

2 The Problem of Immanent Comparisons

The "Problem of Immanent Comparisons" (PIC) begins with an appeal in Cohen and Callender (2009) to a distinction in Quine between *immanent* and *transcendent* notions. Quine writes: "A notion is immanent when defined for a particular language; transcendent when directed to languages generally" (Quine 1970, p. 19). Measurements of simplicity, since they depend on the language in which a system is expressed, are taken by Cohen and Callender to be immanent in this Quinean sense. Strength, or informativeness, is similarly immanent, since it is assumed to depend on the expressive power of the language in which a system is expressed. And, to finish out the set, balance

is said to be immanent as well, since it will be a measure dependent on immanent measures of simplicity and strength. If two systems are competing to be the best and are expressed in different languages, then we would need transcendent measures of simplicity, strength, and balance, in order to implement the best system competition. But "there are too few (viz. no) transcendent measures" of simplicity, strength, and balance (Cohen and Callender 2009, p. 8). Cohen and Callender write that

> *Prima facie*, the realization that simplicity, strength, and balance are immanent rather than transcendent—what we'll call *the problem of immanent comparisons*—is a devastating blow to the [BSA and its variants]. For what counts as a law according to that view depends on what is a Best System; but the immanence of simplicity and strength undercut the possibility of intersystem comparisons, and therefore the very idea of something's being a Best System.
>
> (Cohen and Callender 2009, p. 6, emphasis in original)

The only solution to the PIC, since (supposedly) systems can only be compared when they are expressed in the same language, is to adopt single language privileging.

3 Neither Immanent nor Transcendent

The issue with the PIC is that it ignores the existence of a large middle ground of measures that are neither immanent nor transcendent. To start, let us examine the central claim of the PIC: that simplicity, strength, and balance must be immanent measures. In defense of the idea that simplicity is immanent, Cohen and Callender

(2009, p. 5) defer to Goodman (1954) by way of Loewer, who writes: "Simplicity, being partly syntactical, is sensitive to the language in which a theory is formulated" (Loewer 1996, p. 109). Loewer and Goodman are exactly right. Simplicity is language sensitive. For example, let us adopt a naive version of simplicity, $SimpC(-)$, that is measured by the number of characters it takes to express a sentence (including spaces and punctuation). Consider the following sentence.

This sentence is simple.

Its *SimpC*-simplicity is 24 characters. The same sentence in Dutch is

Deze zin is eenvoudig.

The sentence's *SimpC*-simplicity now is 22 characters. So the *SimpC*-simplicity of a sentence depends or is sensitive to the language in which the sentence is expressed. Does that language sensitivity mean that *SimpC* is immanent? It depends on what is meant by being "defined for a particular language".

SimpC is, in some sense, "defined for a particular language". Insofar as the measure gives conflicting results for a sentence expressed in different languages, it would be ill-defined if we took it to be directed at sentences irrespective of the language in which they are expressed. One way of dealing with this would be to think that we have a multitude of distinct simplicity measures: $SimpC_{\text{English}}(-)$, $SimpC_{\text{Dutch}}(-)$, and so on. But doing that disguises an important fact: each of these measures of simplicity is *the same measure*, just relativized to particular languages. Drawing our inspiration from the "package deal" of Loewer (2007)—in which the BSA holds its competition between system-language pairs (or packages)—we could just as easily deal with the language

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -128-

sensitivity of *SimpC* by saying it is defined for sentence-language pairs. We don't need, then, different measures of simplicity. Just the one will do:

$$\begin{aligned} SimpC(\ulcorner\text{This sentence is simple.}\urcorner, \text{English}) &= 24 \text{ char.} \\ SimpC(\ulcorner\text{This sentence is simple.}\urcorner, \text{Dutch}) &= 22 \text{ char.} \end{aligned}$$

In this way, *SimpC* is better understood as transcendent, and not immanent, because it is, as Quine put it, "directed to languages generally".

Of course, *SimpC* can't be directed to *all* languages, since it will be undefined for any languages that don't have a written form with discrete characters. This suggest that there is an important middle ground between immanent and transcendent measures. When a measure falls in that middle, as *SimpC* seems to, I will say that it is a "moderate measure".

So which conception of *SimpC* is the right one? The "devastating blow" that immanence deals to the BSA and its variants is that it "undercut[s] the possibility of intersystem comparisons" (Cohen and Callender 2009, p. 6). In our naive example,

$$SimpC_{\text{English}}(\ulcorner\text{This sentence is simple.}\urcorner)$$

is—if *SimpC* is immanent—incomparable to

$$SimpC_{\text{Dutch}}(\ulcorner\text{This sentence is simple.}\urcorner).$$

But obviously it's not. ⌜This sentence is simple.⌝ is *SimpC*-simpler in Dutch than in English (when being *SimpC*-simpler means having a lower value of *SimpC*).

Nothing prevents a transcendent or moderate measure from taking a language as one of its arguments. Such a measure is transcendent (or moderate), but language sensitive, and, importantly, it allows for comparisons even when a variety of languages are involved. That being the case, the mere language sensitivity of simplicity, strength, and their balance is not enough to guarantee that they are immanent, nor is it enough to guarantee the incomparability of systems expressed in different languages.

In response to the existence of a measure like *SimpC*, it might be suggested that there may well be transcendent (or moderate) measures plausibly named "simplicity" (etc.), but these are not the ones relevant to the BSA; the measures that *do* appear in BSA will be immanent. It is absolutely right to question the plausibility of a measure as naive as *SimpC* having a role to play in the BSA. (I certainly do not intend to defend *SimpC* as the right measure of simplicity for the BSA.) But I do not think it is clear why we should assume that the right measures are immanent. Rather, I think that moderate measures are, if anything, the norm, and an example may be found in the selection of statistical models.

Following Forster and Sober (1994), statistical model selection has standardly been associated in philosophy with the Akaike Information Criterion (AIC):

$$AIC(M) \;=\; 2[\text{number of parameters of } M] - 2[\text{maximum log-likelihood of } M]$$

The full details of AIC are not terribly important for our purposes here; it is enough to point out that that first term is concerned with the *number of parameters* of the statistical model M. Forster and Sober note that the number of parameters "is not a merely linguistic feature" of models Forster and Sober (1994, p. 9, fn. 13). But the

number of parameters is *a* linguistic feature of a model. Since AIC can compare models with different numbers of parameters, it can—if we think of statistical models as the system-language pairs of the BSA, and AIC as central to the best system competition[2]—compare systems expressed in different languages. AIC is thus a moderate measure.

It is important to note, however, that AIC is also not a transcendent measure. Kieseppä (2001) offers a response to critics of AIC who are concerned that the measure is sensitive to changing the number of parameters of a model by changing the model's linguistic representation. The response turns on the justification of "Rule-AIC", which says to pick the model with the smallest value of AIC, on the grounds that the predictive accuracy of model M is approximately the expected value of the maximum log-likelihood of M minus the number of parameters of M. Crucially,

> the theoretical justification of using (Rule-AIC) is valid when the considered models are such that the approximation [just mentioned] is a good one.
>
> (Kieseppä 2001, p. 775)

Let M be parameterized to have either k or k' parameters. Then there are two claims that are relevant to the justification of Rule-AIC:

$$\text{predictive accuracy of } M \approx E[(\text{maximum log-likelihood of } M) - k]$$

$$\text{predictive accuracy of } M \approx E[(\text{maximum log-likelihood of } M) - k']$$

[2]To make the connection between AIC and the BSA even stronger, it it worth noting that Forster and Sober (1994) take the "number of parameters" term to be tracking the simplicity of a model.

Seattle, WA; 1-4 November 2018 -131-

The predictive accuracy of M is independent of the number of parameters used to express M.[3] But the right side of the approximation in each claim *does* depend on the number of parameters. In general, both of these claims will not be true. Since Rule-AIC is only justified by the truth of these approximations, it will only be applicable to whichever parameterization of M makes the approximation true. The only time when both claims are true, and thus when AIC is applicable to both parameterizations, is when the difference between $E[(\text{maximum log-likelihood of } M) - k]$ and $E[(\text{maximum log-likelihood of } M) - k']$ is negligible. Kieseppä concludes:

> This simple argument shows once and for all that the fact that the number of the parameters of a model can be changed with a reparameterisation does not in any interesting sense make the results yielded by (Rule-AIC) dependent on the linguistic representation of the considered models.
>
> (Kieseppä 2001, p. 776)

From the epistemic perspective that is Kieseppä's concern, I can find room to agree that there is no "interesting sense" in which Rule-AIC is language dependent. This is because, if we are looking to employ Rule-AIC in statistical model selection, what is available to us is a procedure to check if the given parameterization is one that can support the justification of Rule-AIC. If the justification will work, then Rule-AIC applies, and if not, not. Rule-AIC isn't language dependent "in any interesting sense" insofar as it simply doesn't apply to the problematic languages/parameterizations that undermine its justification.

[3]This is intuitively true. It is also true in the formal definition of predictive accuracy given in Kieseppä (1997) and used in this argument from Kieseppä (2001).

However, from the perspective of the BSA and the PIC, these failures of Rule-AIC *are* interesting. AIC (the measure) is not immanent, but it is also not transcendent; it is merely moderate. *Some* reparameterizations of considered models will lead to the inapplicability of Rule-AIC. If Rule-AIC was how we were deciding which system was best, the existence of these problematic reparamterizations would be, as Cohen and Callender put it, a *prima facie* devastating blow to the BSA.

Towards the end of their introducing the PIC, Cohen and Callender write that

> What is needed to solve the problem is a *transcendent* simplicity/strength/balance comparison of each axiomatization against others. The problem is not that there are too many immanent measures and nothing to choose between them, but that there are too few (viz., no) transcendent measures.
>
> (Cohen and Callender 2009, p. 8, emphasis in original)

Cohen and Callender are probably right that there are "too few (viz., no) transcendent measures". In response to this, PIC says that measuring the goodness of a system must be done with immanent measures, and so no systems expressed in different languages may be compared in the best system competition. But non-transcendence is not a guarantee of immanence. We might call the problem that remains the *problem of transcendent measures* (PTC). Measures like AIC are not immanent, but they also aren't transcendent. That non-transcendence gives rise to a degree of language sensitivity that will *sometimes* prevent us from comparing systems expressed in different languages.

In response to the PIC and the supposed immanence of measures appropriate for the BSA, Cohen and Callender (2009) proposed the Better Best Systems Analysis (BBSA),

which relativizes laws to single languages. According to the BBSA, a best system competition is run for every language L (with some restrictions on "every" that aren't especially important here) where all the competing systems are expressed in L and the theorems of the system that is the victor of the competition are the laws *relative to* L. But now it seems that we might have at our and the BSA's disposal moderate measures. In the face of the non-transcendence of these measures—that is, in the face of the PTC—the BBSA's strategy of language relativity is still a good one.[4] Our language relativity does not, however, have to involve privileging *single* languages. The alternative is to relativize to *classes* of languages constructed to ensure the applicability of the measures employed in our best system competition.

4 Discovering Laws and Kinds Together

Before saying more about what relativizing laws to classes of languages would be like in any detail, it is important to say something about why we should pursue language-class relativity over the single language relativity of the BBSA. So, why should we? The reason is that one of the great virtues of the BSA and its variants is their offering of a metaphysics for laws that parallels the search for laws that is to be found in scientific practice, and that parallel is broken by single language privileging. A feature of the

[4]Without going into excessive detail about benefits (and costs) of the BBSA's relativity strategy over competitors, I hope it is enough to note that relativizing the laws allows us to sidestep the question of which language should be privileged entirely, since, ultimately, all languages will get a turn at being privileged, and thus, effectively, none are privileged over all.

search for laws in scientific practice is that it happens in conjunction with a search for the basic kinds of the world. This feature encourages us to acknowledge the importance of language in the BSA, since the basic kinds of the world are, presumably, going to correspond with the basic kinds that appear in the language in which the laws are expressed. Thus, when Lewis first recognizes the language sensitivity of simplicity, he concludes on a celebratory note by saying that the variant of single language privileging he introduces has the virtue of "explaining" why "laws and natural properties get discovered together" (Lewis 1983, p. 368).

For Loewer's Package Deal Analysis, the idea that laws and kinds are discovered together is central to the view. Indeed, the phrase "package deal" has its roots in Lewis, who says just before the "discovered together" remark that "the scientific investigation of laws and of natural properties is a package deal" (Lewis 1983, p. 368). While Loewer ultimately endorses a version of single language privileging, it is accompanied with a rough account of how a "final theory"—i.e., a candidate system-language pair—is arrived at:

> a final theory is evaluated with respect to, among the other virtues, the extent to which it is informative and explanatory about truths of scientific interest as formulated in [the present language of science] SL or any language $SL+$ that may succeed SL in the rational development of the sciences. By 'rational development' I mean developments that are considered within the scientific community to increase the simplicity, coherence, informativeness, explanatoriness, and other scientific virtues of a theory.
>
> (Loewer 2007, p. 325)

If the practice of science parallels the Package Deal Analysis, then the processes of discovering the laws and basic kinds are one and the same.

And it seems Cohen and Callender are also on board with laws and kinds being discovered together when they offer this nice remark on the phenomenon:

> historical disputes between theorists favoring very different choices of kinds seem to us to be disputes between two different sets of laws [...] it has happened in the history of science that people have objected to particular carvings—most famously, consider the outrage inspired by Newton's category of gravity. But given the link between laws and kinds, this outrage is probably best seen as an expression of the view that another System is Best, one without the offending category. If that other system doesn't in fact fare so well in the best system competition—as in the case of the systems proposed by Newton's foes—then the predictive strength and explanatory power of a putative Best System typically will win people over to the categorization employed. While it's true that some choices of [kinds] may strike us as odd, no one would accuse science—the enterprise that gives us entropy, dark energy, and charm—as conforming to pre-theoretic intuitions about the natural kinds of the world. Yet these odd kinds are all embedded in systematizations that would produce what we would consider laws.
>
> (Cohen and Callender 2009, pp. 17–18)

With everyone in agreement, what is the problem? Language privileging, essentially, happens *before* the identification (in the BSA and its variants) or discovery (in scientific practice) of the laws. Though Cohen and Callender will not "accuse science" of

"conforming to pre-theoretic intuitions about the natural kinds of the world", that is exactly what the BBSA (and any other single language privileging variant of the BSA) does when it privileges sets of kinds prior to a best system competition. Furthermore, PIC makes it such that "the predictive strength and explanatory power of a putative Best System" cannot "win people over to the categorization employed" because comparing two putative Best Systems expressed in different languages (with different "categorizations") is supposed to be impossible.[5]

Relativizing to classes of languages solves this problem. Scientists are able to approach the discovery of laws and kinds with pre-theoretic intuitions about how to systematize the world, the language to use when doing that, and the best system competition. As we will see below, the intuitions regarding language and the best system competition will locate them in a particular language class. Scientists will move away from their intuitions about language (and systematizing) when, much as Loewer describes above, there are languages in the relevant language class that may be paired with systems to yield a system-language pair that is scored better by the best system competition than the pre-theoretic system-language pair.[6]

[5]At least, it is impossible according to PIC for the BSA and its variants. If it *is* possible for scientists, then it is wholly unclear why it would be impossible for the BSA.

[6]This movement is only metaphorical for the BSA, where all the possibilities are considered and judged simultaneously. It is helpful, though, to think in the more methodical terms—of considering particular transitions from one system-language pair to another, the benefits that they might bring, and then adopting them or not—because that is what will happen in actual scientific practice.

5 Limiting Language Relativity

Let us begin addressing how language-class relativity can work by looking in more detail at the single language relativity of the BBSA. In the BBSA, there are the fundamental kinds K_{fund}. The set of all kinds $\mathcal{K}$ is the set including K_{fund} closed with respect to supervenience relations—that is, $\mathcal{K}$ includes every kind that can be defined as supervening on the arrangement of the K_{fund} kinds in the actual world. A language L is determined by the set of kinds for which it has basic predicates, and there is a language L_i for every $K_i \subseteq \mathcal{K}$. For any two languages L and L', the supervenience relations between the kinds of the languages and K_{fund} can be thought of as schemes for *translation* between L and L'. The set of all languages $\mathcal{L}_{\text{all}}$ can be thought of as the set of languages that includes L_{fund} closed with respect to all translations. A class of languages $\mathcal{L}_i$ is a set of languages including L_{fund} closed with respect to some acceptable (all, in the case of $\mathcal{L}_{\text{all}}$) translations.

To illustrate, let us consider a 'coin flip' world. Such a world is a string of `H`s and `T`s, which we will assume are the only two fundamental kinds. Another set of kinds might be $K_{\text{ex}} = \{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}\}$, where the translation that gets us to the corresponding language L_{ex} from L_{fund} maps the pairs `HH`, `HT`, `TH`, and `TT`, to `a` through `d`, respectively. An example of a class of languages that includes L_{ex} could be $\mathcal{L}_{\text{n-tuple}}$: Let an acceptable translation for $\mathcal{L}_{n\text{-tuple}}$ be one that, for a given n takes the set of all n-tuples of `H` and `T`, and maps them to a set of kinds $K_n = \{k_{n,1}, k_{n,2}, ...k_{n,2^n}\}$. L_{fund}, then, is just L_1. When `a` through `d` are $k_{2,1}$ through $k_{2,4}$, our K_{ex} and L_{ex} are precisely K_2 and L_2. All, and only, the languages that may be formed through this procedure will be members of the class $\mathcal{L}_{\text{n-tuple}}$.

A language-class relative variant of the BSA will run a best system competition for

every class of languages $\mathcal{L}_i$. Then $\mathcal{S}$ is the set of all systematizations of the world, the set of all competing system-language pairs for the $\mathcal{L}_i$-relative best system competition is given by $\mathcal{S} \times \mathcal{L}_i$.

We can apply this conception of language-class relativity to our other running example of statistical model selection with AIC. Recall that *some* reparameterizations of statistical models would prove problematic for the use of AIC. To reparameterize a model is akin to translating it from one language to another. We can understand, then, the problem of language sensitivity for AIC as being related to some set of problematic translations. If we subtract these problematic translations from the set of all translations, then we have a set of acceptable translations which defines a class of languages that we can call $\mathcal{L}_{AIC}$. $\mathcal{L}_{AIC}$ is precisely the set of all languages such that a system expressed in any one of them will be comparable to a system expressed in any other using AIC. As long as the moderate measures used in the best system competition have clearly problematic and/or acceptable translations associated with them, then the class of languages that may be used to express competing systems will be determined by the measures used in the best system competition.

This will have one of two effects on the extent to which the BSA must be relativized to classes of languages, but before going into those details it will be helpful to characterize "competition relativity". Competition relativity should be understood in much the same way that language relativity is understood. The competition of the BSA is the thing that takes system-language pairs as its inputs, and outputs a best pair from which we can read off the laws. The competition decides what system-language pair is best by considering how well they measure up with respect to some collection of theoretical virtues (like simplicity and informativeness) and the actual world. Much as

we might worry about what language to privilege, and side-step that problem by relativizing laws to languages so that every language takes a turn as the privileged one, we might also worry about which competition, or which set of theoretical virtues, to privilege. Competition relativity sidesteps the problem of which collection of theoretical virtues to use (and weighting between them, and means of measuring them, etc.) by relativizing laws to every way of formulating a best system competition.[7]

So, either the BSA will be committed to competition relativity or not. Suppose that it is not. For convenience, suppose further that Rule-AIC is all that there is to the best system competition. In that case, the BSA will always be run using the $\mathcal{L}_{AIC}$ class of languages. Language-class relativity is not required since there is only one language class that will ever be relevant to the BSA—namely $\mathcal{L}_{AIC}$, as determined by the best system competition. Now suppose that there is competition relativity. A different best system competition must be run for every competition function C_i in the set of all possible competition functions $\mathcal{C}$. In principle we will need to run best systems competitions for every pair in $\mathcal{C} \times \mathbb{L}$, where $\mathbb{L}$ is the set of all language classes. Let $\mathcal{L}_j$ be the class of languages constructed according to the translations that are acceptable for the measures that comprise C_i when $i = j$. In practice, however, it will only make sense to run a competition once for each $C_i \in \mathcal{C}$, since the pairs $C_i, \mathcal{L}_j$ will be unproblematic only when $i = j$. Language-class relativity in this situation will be redundant with competition relativity. We also have it that, in either case (of needing competition relativity or not), single language relativity remains unnecessary for all the same reasons that recommended language-class relativity.

[7]See Bialek (2017) for an extended discussion of competition relativity and the possibility of its inclusion in the BSA.

6 The Trivial Systems Problem

The redundancy of any sort of language privileging relativity with competition relativity offers an interesting solution to the Trivial Systems Problem (TSP) that initiated the trend of single language privileging.

Recall that the TSP is concerned with the possibility of suitably gerrymandered languages that can guarantee that the "best" system will have axioms and theorems undeserving of the name "law". In the introduction fo the problem, Lewis imagines a system S and predicate F "that applies to all and only things at worlds where S holds" (Lewis 1983, p. 367). The system S, then, maybe be expressed by the single axiom $\forall xFx$, simultaneously achieving incredible informativeness—because of the specific applicability of F—and incredible simplicity—because, Lewis assumes, '$\forall xFx$' is about as simple as a system could be. So S will be the best system despite a variety of reasons why it shouldn't be, the foremost of which are that: (1) $\forall xFx$ will be a law unlike any we would expect to find, (2) F would be a basic kind unlike any we would expect to find, and (3) every regularity of the world is a theorem of $\forall xFx$, so there would be no distinction between accidental and lawful regularities.

The problem is solved as long as we can avoid languages that include problematic predicates like F. Single language privileging solves this problem as long as the privileged language does not include the (or any) problematic predicate(s). Language-class privileging likewise solves the problem as long as no language in the class includes the (or any) problematic predicate(s). That alone might be enough said, but the redundancy of language-class choice on competition choice offers a more nuanced solution: The best system competition could be chosen such that the corresponding class

of languages does not include F or any similarly problematic predicates. But it could also be chosen such that F and its ilk are certain to not be the best. Lewis assumes with no discussion that $\forall xFx$ is an incredibly informative and simple system, but, even if that is true for the measures/competition, it need not be true for every competition. If there is competition relativity, then there may be competitions for which a trivial system like $\forall xFx$ is the victor, but for the same reasons that such a system is problematic, scientists will simply be uninterested in the laws relative to those competitions.[8] If there isn't competition relativity, it seems unlikely that science would unequivocally endorse a competition that yields a trivial system (or, if it does, then we would need to take a step back and seriously reconsider our aversion to such a system).

In the end, there is no apparent need for any language privileging or relativity in the BSA.[9] Its role in solving the problems of immanent (or transcendent) comparisons and trivial systems will be unnecessary (if a single moderate best system competition can be identified) or redundant with competition relativity.

[8]In much the same way that Cohen and Callender (2009) allow for there to be uninteresting sets of laws determined relative to languages that include F-like predicates.

[9]The problems discussed is not the only reason one might want to adopt language relativity in the BSA. It should also be noted that one of the virtues of the BBSA's single language relativity is that it allows the view to accommodate an egalitarian conception of special science laws. Language relativity, however, is not the only way of getting special science laws out of the BSA. This is an important issue to which the discussion in this paper is relevant, but a proper exploration of it warrants a more focused and extended treatment.

References

Bialek, M. (2017). Interest relativism in the best system analysis of laws. *Synthese 194*(12), 4643–4655.

Cohen, J. and C. Callender (2009). A better best system account of lawhood. *Philosophical Studies 145*(1), 1–34.

Forster, M. and E. Sober (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science 45*(1), 1–35.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Kieseppä, I. (1997). Akaike information criterion, curve-fitting, and the philosophical problem of simplicity. *The British journal for the philosophy of science 48*(1), 21–48.

Kieseppä, I. (2001). Statistical model selection criteria and the philosophical problem of underdetermination. *The British journal for the philosophy of science 52*(4), 761–794.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Loewer, B. (1996). Humean supervenience. *Philosophical Topics 24*(1), 101–127.

Loewer, B. (2007). Laws and natural properties. *Philosophical Topics 35*(1/2), 313–328.

Quine, W. V. O. (1970). *Philosophy of logic*. Harvard University Press.

Schrenk, M. (2008). A theory for special science laws. In S. W. H. Bohse, K. Dreimann (Ed.), *Selected Papers Contributed to the Sections of GAP.6*, pp. 121–131. Paderborn: Mentis.

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.

Explaining Scientific Collaboration: a General Functional Account

Thomas Boyer-Kassem* and Cyrille Imbert†

October, 2018

Abstract

For two centuries, collaborative research has become increasingly widespread. Various explanations of this trend have been proposed. Here, we offer a novel functional explanation of it. It differs from accounts like that of Wray (2002) by the precise socio-epistemic mechanism that grounds the beneficialness of collaboration. Boyer-Kassem and Imbert (2015) show how minor differences in the step-efficiency of collaborative groups can make them much more successful in particular configurations. We investigate this model further, derive robust social patterns concerning the general successfulness of collaborative groups, and argue that these patterns can be used to defend a general functional account.

*MAPP (EA 2626), Univ. Poitiers, France. thomas.boyer.kassem@univ-poitiers.fr
†CNRS, Archives Poincaré, France. cyrille.imbert@univ-lorraine.fr

1 Introduction

For two centuries, co-authoring papers has become increasingly widespread in academia (Price, 1963, Beaver and Rosen, 1979), especially in the last few decades. Since the 1950s, the percentage of co-authored papers has grown at a common rhythm for science and engineering, social sciences, and patents; the mean size of collaborative teams has also increased, and even more so in science and engineering. No such increase is visible for the art and humanities (Wuchty et alii, 2007).

Various explanations of this collaborative trend have been proposed: for example, it may be caused by scientific specialization, it may increase the productivity or reliability of researchers, or be promoted by the rules of credit attribution. Here, we aim at offering a new functional explanation of this trend by showing that collaboration exists because it increases the successfulness of scientists. The present explanation differs from accounts like that of Wray (2002) by the social and epistemic mechanism that grounds the beneficialness of collaboration. We analyze further an existing model that shows how minor differences in the step-efficiency of collaborative groups at passing the steps of a project can make them much more successful in particular configurations (Boyer-Kassem and Imbert, 2015) and show how it can be used to build a general and robust functional explanation of collaboration.

We introduce the model in section 2. After presenting functional explanations (section 3), we show how the model can be used to derive robust social patterns of the successfulness of collaborative groups (section 4), and argue that these patterns can refine and strengthen functional explanations of collaboration like the one defended by Wray (sections 5 and 6).

2 Boyer-Kassem and Imbert's Model: Main Results and Explanatory Lacunas

Boyer-Kassem and Imbert (2015) investigate a model in which n agents struggle over the completion of a research project composed of l sequential steps. At each time interval, agents have independent probabilities p of passing a step. When an agent reaches the end of the project, she wins all the scientific credit and the race stops (this is the priority rule). Agents can organize themselves into collaborative groups for the whole project, meaning that they only share information, i.e. step discoveries — clearly, there are more favorable hypotheses associated with collaborating, like having new ideas or double-checking (see below). Within a group, agents make progress together, and equally share final rewards. Thus, a group of k agents (hereafter k-group) passes a step with probability $p_g(k,p) = 1-(1-p)^k$. In forthcoming illustra-

tions, the value of l is set to 10 and that of p to 0.5, which is not particularly favorable for groups (ibidem, 674). If collaboration is beneficial with these hypotheses, it will be even more so with more favorable or realistic ones. A community of n agents (hereafter, n-community) can be organized in various k-groups. For example, a 3-community can correspond to configurations (1-1-1), (2-1) or (3). The individual successfulness of an agent in a k-group in a particular configuration is defined as the average individual reward divided by time. It has been obtained for all configurations up to $n = 10$, on millions of runs.

Note that this model is not aimed at quantifying the actual successfulness of collaborative agents, but at analyzing the differential successfulness of agents depending on their collaborative behavior. The main finding is that minor differences in the efficiency at passing steps can be much amplified and that, even with not-so-favorable hypotheses, collaboration can be extremely beneficial for scientists. For example, in a (5-4) (resp. (2-1)) configuration, whereas the difference in step efficiency between the 5 (resp. 2) and the 4-group (resp. 1-group) is 3% (resp. 50%), the difference in individual successfulness is 25% (resp. 700%). The scope of these results actually goes beyond the initial hypotheses in terms of information sharing. Formally speaking, the model is a race between (collective) agents i with probabilities p_i of passing steps. *Whatever the origin* of the differences in p_i, they are greatly amplified by the sequential race. In other words, any factor, whether epistemic or not, that implies an increase in p_i of a k-group (e.g. if a collaborator is an expert concerning specific steps, if increased resources improve step-efficiency, etc.) makes this group as successful as a larger group — hence the generality of this mechanism.

Still, these results do not explain scientific collaboration by themselves. First, collaboration is beneficial for particular k-groups in particular configurations only: a 2-group is very successful in configuration (2-1-1-1-1) but not in (7-2). Thus, the model mostly provides possibility results about what can be the case in certain configurations. Second, the explanandum is a general social feature of modern science, not some collaborative behavior in some particular case, so the explanans must also involve general statements about the link between collaboration and beneficialness. Then, if the model presents generic social mechanisms with explanatory import, one needs to describe at a general level the effects of these mechanisms and provide some general, invariant pattern between collaboration and beneficialness. This is what we do in section 4. A final serious worry is that the beneficialness of a state by no means explains why it exists, nor perseveres in being. A link needs to be made between the beneficialness of collaboration and its existence over time. We suggest that this connection can be accounted for functionally.

3 Functional Explanations and Collaboration

We review in this section how functional explanations work and how they can be used in the present case. We follow Wray's choice to use Kincaid's account because it is simple, widely accepted, and that nothing substantial hinges on this choice. Functional explanations explain the existence of a feature by one of its effects, usually its usefulness or beneficialness. As such, they can be sloppy and badly flawed. The usefulness of the nose to carry glasses does not explain that humans have one. Nevertheless, if stringent conditions are met, it is usually considered that functional explanations can be satisfactory, typically within biology. Even Elster, who otherwise favors methodological individualism, agrees that functional explanations can be acceptable in the social science (Elster, 1983). According to Kincaid (1996, 105-114), P is functionally explained by E, i.e. P exists "in order to promote <effect E>" if:

(1) P causes E,

(2) P persists because it causes E,

(3) P is causally prior to E.

Then, a functional explanation of collaboration should have the following form:

(1c) Scientists' collaborative behavior causes the increase of their individual successfulness.

(2c) Scientists' collaborative behavior persists (or develops) because it causes a higher individual successfulness.

(3c) Collaborative behavior is causally prior to this increased individual successfulness that is rooted in collaborative behavior.

We agree with Wray (2002, 161) that it is implausible to consider that the high successfulness of scientists is the initial cause of collaboration since many scientists have been successful (and continue to be in some fields) without collaborating. In the same time, there can be various contingent reasons why some researchers have decided to engage in some collaboration. So, what calls for an explanation is the fact that collaboration is widespread and persistent, not its occasional existence.

4 Collaboration Causes Successfulness

We now argue that the above model provides strong evidence in favor of (1c). To explain the general collaborative patterns described above, the causal

relation between collaboration and successfulness needs to be general and robust. Hence, one needs to go beyond the description of the beneficialness of collaboration in particular situations. A first route is to find general results about when it is beneficial for individuals to collaborate, such as the following theorem (see the appendix for the proof).

Theorem. When m groups of equal size k merge, the individual successfulness of agents increases.

In other words, as soon as several k-groups of the same size exist, they would improve the individual successfulness of their members by merging. A corollary is that single individuals always have interest in collaborating. However, this theorem only covers a small subset of possible configurations, and cannot provide a general vindication for the causality claim (1c). Further, agents might only use it if they are aware of it and are in a position to identify groups of equal-size competitors, which cannot be assumed in general.

To overcome these difficulties, we now assess agents' successfulness irrespective of what they know about other competitors: we consider the average successfulness of k-groups over all possible configurations for each community size. For example, we average the individual successfulness of 4-groups in configurations (4-1-1-1); (4-2-1) and (4-3)[1]. In order to study the robustness of the causal relation between collaboration and successfulness, we investigate in the next paragraphs how much collaborating remains beneficial under variations of key parameters of the competition context.

Successfulness and community size. Figure 1 shows the average successfulness within k-groups for communities of various sizes. First, the successfulness of loners brutally collapses and is much lower than that of other k-groups as soon as $n > 2$. This confirms that except when nobody collaborates, or in very small communities, loners are outraced. Second, for all group sizes, individual successfulness decreases for larger communities, as can be expected when the number of competing groups and their size increases. Nevertheless, the successfulness of k-groups remains high and stable up to some community size s larger than k till they are eventually outperformed by larger groups or till growing bigger would mean over-collaborating (see (Boyer-Kassem and Imbert, 2015, 679-80) for an analysis of over-collaboration in large groups). Third, the larger the groups are, the longer and flatter this initial plate of successfulness is and the less steep the decrease in successfulness is. Fourth,

[1]There is no clear rationale about how to weigh configurations. From a combinatorial viewpoint, configuration (1,1,1,1,1,1,1) has one realization and (3,2,1,1) several ones. But from an empirical viewpoint, when scientists hardly collaborate, configuration (1,1,1,1,1,1,1) is usual and (3,2,1,1) extremely rare. We have privileged simplicity and chosen to give equal weight to all configurations.

Seattle, WA; 1-4 November 2018 -149-

when n is much larger than k, the successfulness of k-groups increases with k. However, this increase is a moderate one and small groups still do reasonably well, which is somewhat unexpected, given the general amplification effect — but see the analysis of figure 3 below for more refined analyses. Typically, in 10-communities, 2-groups do badly but remain somewhat viable since their average successfulness remains between 1/3 to 1/2 of that of 3 or 4-groups. Overall, not collaborating is in general not a viable strategy. Collaborating moderately ($k = 2$ or 3) can be very rewarding when there are few competitors (e.g. in small research communities, or on ground-breaking questions that are only known to a handful of scientits). Smalls groups remain viable but tend to be outraced when communities become significantly larger (typically, concerning questions belonging to normal science that many researchers are likely to tackle). Thus, moderately collaborating is a viable but more risky strategy when uncertainty prevails about the number and size of competing groups. Finally, while large collaborative groups rarely get exceptionally high gains, they are extremely safe, with moderate differences in successfulness between them or when the community size increases.

Successfulness and group size. Figure 2 shows the variation of individual successfulness with group size for various community sizes. First, for $n > 2$, the successfulness curve has a one-peaked (discrete) form, the maximum of which grows with the community size. Second, these one-peaked curves are not symmetric: the increase in successfulness is steep (but less so for larger groups), the decrease is gradual (idem). Large groups predate resources so groups need to grow big quickly to get some share and because returns can be increasing (Boyer-Kassem and Imbert 2015, 678), the increase in successfulness is steep. The decrease after the peak is slow because large group are hard to predate but over-collaborating can become suboptimal when the increase in gain by predation no longer makes up for the need to share between more people). These results are not trivial because at the configuration level, the successfulness of groups is contextual. They are important, too. A one-peaked profile is usually *assumed* in the literature about coalitions. Here, it emerges from a micro-model, and gets its justification from it. Overall, these patterns show again that agents have a large incentive to collaborate substantially, whatever the competing environment.

Successfulness in more or less collaborative communities. Figure 3 finally shows how the successfulness of k-group members varies with the degree of collaboration in their competition environment.[2] Here again, what matters

[2]Here, the degree of collaboration in each configuration is assessed by computing the average size of k-groups. For each k, we then compute the average successfulness of a member of a k-group over configurations having a degree of collaboration within intervals $[1, 1.5]$ (represented at coordinate "1.25" on the x-axis), $[1.25, 1.75]$, $[1.5, 2]$ ... $[3.5, 4]$. We

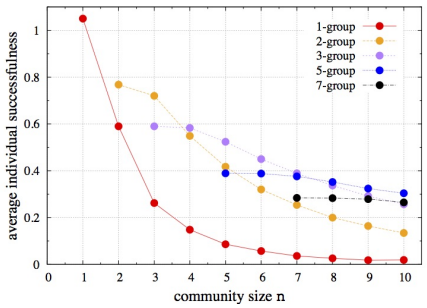


Figure 1: Variation of individual successfulness with community size.

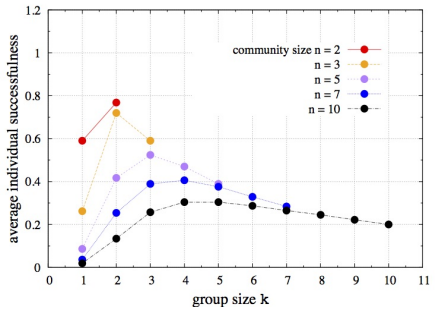


Figure 2: Variation of individual successfulness with the size of groups.

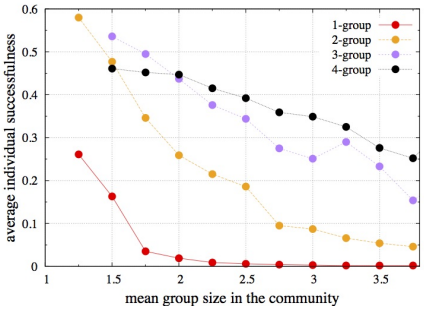


Figure 3: Variation of successfulness with the degree of collaboration in communities.

is less the exact value of the successfulness than the differential successfulness between more or less collaborating individuals. The graph confirms that successfulness depends less on the absolute size of groups than on how much they collaborate in comparison with their competitors. Scientists who collaborate more than average are very successful; those who collaborate as their peers do reasonably well; those that collaborate less than average are outraced by a large margin. This general result is not unexpected given all the above results, but the graph highlights that success for intensively collaborating scientists, and underachievement for under-collaborators can be very large. This is an important finding because if, as we shall see, successful scientists pass over their collaborative habits more than their peers, then the feedback loop provides a mechanism that favors the *increase* of the degree of collaboration by promoting those that collaborate more than others.

have chosen overlapping intervals to smoothen results. The average is computed up to communities of size 10.

Partial conclusion. Overall, the results show that — everything else being equal — collaborating a lot entails successfulness. This relation is robust under changes in the size of communities or in the exact size of groups. Further, those who collaborate more than average are much more successful. Collaborating too much is not a significant problem, under-collaborating is. So, collaborating a lot is a safe working habit, especially in the absence of information about the size and structure of the competing community. In light of this evidence, (1c) seems adequately supported.

5 Collaborative Practices Develop Because of the Success of Collaborative Scientists

We have so far argued that collaborative scientists, especially when they collaborate more than others, are more successful. We now need to argue that, because of this differential successfulness, collaborative habits persist and possibly develop in scientific communities (2c). A wide variety of social mechanisms across scientific contexts can contribute to this feedback loop. Accordingly, we shall be content with giving various evidence that strongly suggests that this link is a likely one.

Transmission. Knowing how and when to collaborate is not straightforward. Like other know-how skills, it can be developed by exercising it with people who already possess the relevant procedural knowledge. In this case, people who already collaborate can endorse this role of cultural transmission for colleagues and above all students (Thagard, 2006). Working with students is an efficient way to train them as scientists (Thagard, 1997, 248—50), so scientists have incentives to enroll students in their collaborative groups. Then, the cultural transmission of collaborative practice does not require any particular effort on top of that. The very circumstances that make collaboration possible and beneficial also make its transmission easier: when a research project can be divided into well-defined tasks, the solutions of which can be publicly assessed and shared, it is easier to enroll other people and thereby transmit collaborative skills to them (ibidem). Thus, collaborative habits can be passed over and need not be reinvented by newcomers.

Transmission opportunities. We now argue that collaborative scientists, because they are more successful, will more often be in a position to transmit their collaborative habits and that the collaboration rate will therefore increase. Within applied science, in which collaboration is also widespread (Wuchty, 2007), research projects are usually directed at finding profitable applications, which can be patented. Thus, fund providers are directly and strongly interested in hiring and providing resource to successful scientists,

who develop such applications. Within pure science, the connection is less straightforward. But because scientific success is the official goal of science, successful scientists can be expected to stand better chances to get good positions and grants, develop research programs, and pass over their collaborative habits.

Note that it is merely needed that the function between the pragmatic rewards of scientists and their success is on average increasing. This remains compatible with the fact that *some* epistemically successful scientists get little resource and *some* unsuccessful scientists get a lot — which seems to be the case. Actually, non-epistemic factors may even tend to over-credit successful scientists, and in particular collaborative ones. First, individual successfulness has been assessed in the model with a conservative estimate. It seems that an agent's publication within a k-group is actually more appreciated than just $1/k$ of a single-authored publication. For instance, a large French research institution in medicine officially weighs the citations of a paper with "a factor 1 for first or last author, 0.5 for second or next to last, and 0.25 for all others" (Inserm 2005). Also, a publication within a 10-group will generally be more visible than one single-authored publication, since more people can promote or publicize collective publications and research topics. Second, sociology of science seems to indicate that scientific credit tends to accrue to a subset of scientists who are perceived as extremely successful — this is the Matthew effect (Merton, 1968). Then, to the extent that access to resources increases with scientific credit, successful collaborative scientists can be expected to benefit from this effect and transmit more their working habits. The concentration of credit and resource may further stimulate collaborative behavior with these fortunate scientists.

Other types of mechanisms may contribute to this process, like conscious ones. So far, agents have only been supposed to follow their working habits and sometimes transmit them. But supplementary intentional or imitative processes may also feed this dynamics [3]. Once winners of the scientific race publish co-authored articles, it becomes easy for others to see that successful scientists are highly collaborative ones. (For instance, if agents of a 3-group are 4 times more successful than a single agent, this means that their groups publishes 12 more articles than this agent). Accordingly, the belief that collaborating is beneficial can be acquired as collaborating becomes usual. Furthermore, resources may accrue to scientific institutions that host individually successful scientists, and indirectly to these scientists. Agents in the model can be reinterpreted as teams or collective entities which decide to share results or to combine their expertise to produce collective articles. Then, these institutions

[3]Kincaid mentions that "complex combinations of intentional action, unintended consequences of intentional action, and differential survival of social practices might likewise make these conditions [(1)–(3) in our Section 3] true" (Kincaid 1996, 112).

and their members will be more successful, may attract resource, and will keep developing and transmitting their working habits.

In light of the above discussion, we believe that the causal connection between the success of collaborative scientists and the persistence and development of collaborative practices is highly plausible.

6 Discussion

Good functional explanations should be unambiguous about when the causal mechanisms that they rely on are efficient. In the present case, the following conditions can be emphasized.

First, conditions for the application of the priority rule should be met. In particular, (i) it should be possible to single out problems and to state uncontroversially when they are solved. Second, for the model to apply, (ii) scientific problems should be dividable into subtasks, and (iii) the solutions of these subtasks should be communicable. Finally, the model assumes that (iv) the completion of these subtasks should be sequential, but our conclusions still hold if this condition is relaxed. Indeed, if some subtasks can be tackled in parallel then the project can be completed even more quickly by different agents of a group, and collaboration is even more successful. Conditions (i)-(iii) are somewhat met in the formal and empirical sciences, less so in the social science, and almost not in the humanities. For example, as noted by Thagard (1997, 249), the humanities do not obviously lend themselves to the division of labor and to teacher/apprentice collaborations. Similarly, the importance of interpretative methods and the coexistence of incompatible traditions may prevent consensus on the nature of significant problems and what counts as a solution. This may account for the differences concerning collaborative patterns in these fields.

As mentioned above, different causal pathways may connect the successfulness of collaborative scientists to the persistence and development of collaborative practices. Thus, conditions for the fulfillment of claim (2c) cannot be uniquely specified. But several points are worth mentioning. First, the activity of epistemically successful scientists should be favored by scientific institutions. This can be the case if it is agreed that scientific success, in the form of publications or patents, is valued and promoted. Concerning scientific results that lead to patents, applications and financial gains, this condition is met when public or private funders value such outputs. Concerning pure scientific results, this means that there should be a wide agreement about which results are scientifically good and significant, and there should exist common and accessible publication venues, the value of which is consensual. Again, these conditions are approximately met in the formal and empirical sciences, less so in the social science and, almost not in the humanities in which scholars do not share paradigms, methods or norms about what is scientifically sound

and significant, and cultural and linguistic barriers can restrain the existence of unified communities and common publication venues. Second, in contexts in which researchers and projects are regularly evaluated, especially by agents or institutions who are not in a position to asses the scientific value of their work, the existence of a common standard of success in terms of publications (through simple and calibrated publication indicators) may even more favor researchers who are successful, and therefore the development of collaboration. Finally, when resources are crucial to carry out or facilitate research, snowball effects can favor even more successful scientists, and in particular collaborative ones. This resource accessibility condition, which is central in Wray's explanation, is not in ours. But we agree that in such cases, the functional mechanisms that we describe will be even stronger. In this sense, our account encompasses Wray's. This condition about resources may be another reason for the difference in collaborative behavior between the formal or empirical sciences, the social sciences and the humanities.

7 Conclusion

We have argued that collaborating a lot is overall a safe and success-conducing practice. This conclusion is robust for various sizes of groups, communities and degrees of collaboration; everything being equal, those who collaborate more than average do better. Then, to the extent that the successfulness of researchers gives them more opportunities to transmit their research habits, the development of collaborative practices in communities can be functionally explained. We have further emphasized that the conditions for this functional pattern to work are specifically met in the scientific fields in which collaboration is well-developed. Accordingly, it seems reasonable to consider that this functional mechanism is an important element of the explanation of the development of collaboration in modern science.

The explanation of collaboration is probably a multi-factorial issue. Nevertheless, an asset of our general functional explanation is that it highlights the unexpected force of beneficial aspects of collaborative activities and suggests important roles for contextual factors that are associated with the rise of collaboration. As such, it is general and unifying. For instance, the competition model shows how the division of scientific labor, the use of specialized experts (Muldoon 2017), or the increased reliability of collaborative teams (Fallis 2006, 200) can increase the probability that groups pass research steps and have amplified effects in terms of successfulness. Similarly, factors like the need to access resources to carry out or facilitate research can create a snowball effect that favors epistemically successful (collaborative) researchers (Wray 2002). And factors like the globalization of research or professionalization (Beaver, 1979) can be seen as conditions favoring the application of the priority rule

and scientific competition.

Finally, while nothing in the model provides an internal limit to the growth of collaboration, one can note that there is a wealth of reasons why collaborating groups cannot develop forever. For example, communities are limited in size, spatially distributed, and collaboration is all the more costly as groups are large. The model could be easily modified to integrate factors that limit the success and development of collaboration.

8 Appendix: Proof of the Theorem

Consider first the simple case where the m k-groups don't have other competitors. By symmetry, all groups have the same probability $1/m$ to win the race and get the reward — call this reward r. So, the individual expected reward is $r/(km)$. Suppose now the groups merge and all km agents collaborate. Each of them will receive the same reward, so their expected individual rewards are $r/(km)$ too. However, what matters in the model is not the expected reward, but the successfulness, which is this quantity divided by time. Because within a collaboration agents share all the steps they pass, the larger km-group will be at least as quick, and sometimes more, than all k-groups — more precisely: for a given drawing of all random variables corresponding to attempts to pass the steps, for all agents and temporal intervals, the km-group will move at least as quickly as all k-groups. So the individual successfulness is at least as high when identical groups merge.

Consider now the case where there are other competitors than the m groups. For a given drawing of all random variables, either the winner is one of the m groups, or another competitor. In the former case, the above reasoning can be made again, and the same conclusion holds. In the latter case, there is nothing to lose, and because the km-group is sometimes quicker than the m k-groups, there can be additional cases where it outcompetes the other competitors; then, the individual successfulness increases with the merging. QED.

9 References

Beaver, Donald deB. and Rosen, Richard (1979) "Studies in Scientific Collaboration: Part III", *Scientometrics*, 1(3): 231-245.

Boyer-Kassem, Thomas, and Cyrille Imbert (2015), "Scientific Collaboration: Do Two Heads Need to Be More than Twice Better than One?" Philosophy of Science 82 (4): 667–88.

Elster, Jon (1983), *Explaining Technical Change: A Case Study in the Philosophy of Science*, Studies in Rationality and Social Change, New York: Cambridge University Press.

Fallis, Don (2006), "The Epistemic Costs and Benefits of Collaboration", *Southern Journal of Philosophy* 44 S: 197–208.

INSERM (2005), "Les indicateurs bibliométriques à l'INSERM", `https://www.eva2.inserm.fr/EVA/jsp/Bibliometrie/Doc/Indicateurs/Indicateurs_bibliometriqu Inserm.pdf`

Kincaid, Harold (1996), *Philosophical Foundations of the Social Sciences*, Cambridge University Press.

Merton, Robert K. (1968), "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered", *Science*, 159 (3810): 56–63.

Muldoon, Ryan (2017), "Diversity, Rationality, and the Division of Cognitive Labor", in Boyer-Kassem, T., Mayo-Wilson, C. and Weisberg, M. (eds.), *Scientific Collaboration and Collective Knowledge*, New York: Oxford University Press.

Price, Derek John de Solla (1963), *Little Science, Big Science*, New York, Columbia University Press.

Thagard, Paul (1997), "Collaborative Knowledge", *Nous* 31(2): 242—261.

— (2006), "How to Collaborate: Procedural Knowledge in the Cooperative Development of Science", *The Southern Journal of Philosophy*, XLIV: 177—196.

Wray, K. Brad (2002), "The Epistemic Significance of Collaborative Research", *Philosophy of Science* 69 (1): 150-168.

Wuchty, Stefan, Jones, Benjamin F. and Uzzi, Brian (2007), "The Increasing Dominance of Teams in Production of Knowledge", *Science* 316(5827): 1036-1039.

Individuating Genes as Types or Individuals: Philosophical Implications on Individuality, Kinds, and Gene Concepts

Ruey-Lin Chen
Department of Philosophy
National Chung Cheng University
This paper will be presented at PSA 2018 meeting at Seattle in November

Abstract

"What is a gene?" is an important philosophical question that has been asked over and over. This paper approaches this question by understanding it as the individuation problem of genes, because it implies the problem of identifying genes and identifying a gene presupposes individuating the gene. I argue that there are at least two levels of the individuation of genes. The transgenic technique can individuate "a gene" as an individual while the technique of gene mapping in classical genetics can only individuate "a gene" as a type or a kind. The two levels of individuation involve different techniques, different objects that are individuated, and different references of the term "gene". Based on the two levels of individuation, I discuss important philosophical implications including the relationship between individuality and individuation and that between individuals and kinds in experimental contexts. I also suggest a new gene conception, calling it "the transgenic conception of the gene."

Keywords: gene concept, individuality, individuation, experiment, classical genetics, transgenic technique

1. Introduction: what is a gene and why individuation matters

"What is a gene?" and its related questions have been asked over and over by philosophers, historians, and scientists of biology (Beurton, Falk, and Rheinberger 2000; Carlson 1991; Falk 1986, 2010; Gerstein et al. 2007; Griffths and Stotz 2006, 2013; Kitcher 1982, 1992; Pearson 2006; Stotz and Griffiths 2004; Snyder and Gerstein 2003; Waters 1994, 2007). Those questions are frequently embedded in discussions about the definition of the term "gene" and the gene concept. As a consequence, the phrase "a gene" in this question usually refers to a type of gene. However, should we use "a gene" to refer to an individual gene, i.e., a gene token? Could it in fact be this?

The question "what is a gene" explicitly implies the problem of identifying genes, and identifying a gene presupposes individuating the gene. In what ways are genes individuated and how do scientists individuate them? I call this *the individuation problem of genes*. This paper shall approach the problem from three different but related perspectives.

From the epistemic perspective, a concept of the gene provides at least a working definition, which by nature is a hypothesis, for scientific research. Any hypothesis of the gene may be in error and may be confirmed only by experimentally individuating particular tokens of some gene. From the semantic perspective, according to a Fregean philosophy of language, the concept of reference usually serves for proper names that refer to individuals or particulars. We may extend the concept of reference to general terms (e. g., "humankind" or "gene kind") for the case in which some token of a kind is presented, and so we use a general term to refer to the kind. This means that at least some token of a kind has to be individuated. This semantic perspective presupposes an ontological perspective: the existence of a kind should be presented or demonstrated by the existence of at least a token of the kind. In the case of the gene, the ontological requirement means that we have to individuate a token of some gene kind. All three perspectives indicate the key status of individuation for answering the question of what a gene is.

According to the literature of analytic metaphysics, "individuation" is understood in a metaphysical and an epistemic sense. In the epistemic sense, someone individuating an object "is to 'single out' that object as a distinct object of perception, thought, or linguistic reference." (Lowe 2005: 75) This epistemic sense presupposes the metaphysical sense, in which what 'individuates' an object "is whatever it is that makes it the single object that it is – whatever it is that makes it one object, distinct from others, and the very object that it is as opposed to any other thing." (Lowe 2005: 75) Bueno, Chen, and Fagan (2018) add a practical sense to the term, interpreting

Seattle, WA; 1-4 November 2018 -159-

"individuation" as a practical process through which an individual is produced. They characterize the relation between "individuation" and "individuals" as when "an individual emerges from a process of individuation in the metaphysical sense. Epistemic and practical individuation, then, are processes that aim to uncover stages of that metaphysical process." (Beuno, Chen, and Fagan 2018) The approach to the individuation of genes I adopt herein follows their characterization, especially by focusing on the process of epistemic and practical individuation. Reversely, the case I am investigating in this paper offer an illustration for the new sense of individuation.

Although philosophers have investigated concepts of the gene and its change by examining many cases in scientific practices, they have seldom considered the role that the transgenic technique developed in biotechnology may play in philosophical discussions. This paper explores experimental individuation of genes from the direction of that technique, considering the possibility that a gene is individuated as an individual in the relevant contexts.

This paper thus addresses two central questions: (Q1) In what sense, can we reasonably say that classical geneticists have individuated a gene? (Q2) Are there experiments that can individuate a gene as an individual? Some new questions such as the relationship between individuality and individuation will be derived from the answer to the two questions. This paper is thus structured in the following way.

In the second section, I review the literature about the concepts and references of genes. Section 3 argues that the answer to Q1 is that the geneticists individuate a gene as a type, because they used the chromosomal location technique. Section 4 argues that the answer to Q2 is the experiments that use the transgenic technique. The two answers indicate two different kinds of individuation: individuation of a type and individuation of an individual. This raises a new question about whether or not "individuation of a type" is a consistent phrase. In order to respond to this, section 5 discusses in what sense we individuate a type and compare between two kinds of individuation defined by two different kinds of experiments and techniques: the chromosomal locution of genes and the transgenic experiment. My argument thus involves the relationship between kind and individual in the context of experimentation. Given the new question, Section 6 argues that transgenic experiments can demonstrate a gene type by individuating its tokens, while gene mapping experiments in classical genetics only individuate gene types. Thus, a new gene conception, calling it "the transgenic conception of the gene," can be proposed. I further discuss the relationship among the classical gene concept, the molecular gene concept, and the transgenic conception. In the seventh section, I defend the thesis that practices of individuation in scientific investigations are prior to characteristics of individuality identified by traditionally metaphysical speculations.

2. Concepts and references of the gene

The rapid change of the gene concept has produced a large multitude of gene concepts that have bewildered scientists (Gerstein et. al. 2007; Pearson 2006; Stotz and Griffiths 2004). The confused situation has attracted many philosophers and scientists to provide clarifying analyses. Although scientists as well as philosophers have made endeavors to overcome the predicament, they are motivated differently. Scientists believe that they need a unified concept to help them conduct research and to communicate with each other, because, as developmental geneticist William Gelbert says, "it sometimes [is] very difficult to tell what someone means when they talk about genes because we don't share the same definition" (Pearson 2006: 401). Thus, most scientists seek to redefine the "gene" and tend to adopt a single preferred perspective on the gene concept, although they are well aware with the plurality of gene definitions (Wain et. al. 2002; Gerstein et. al. 2007).

Philosophers at different times have been interested in clarifying concepts of the gene and in investigating the patterns of associated conceptual change. In contrast to actual definitions used by working scientists, they often consider more abstract concepts of the gene that can guide several different definitions in the context of scientific research. Consequently, they conclude that it is almost impossible to find a unified concept of the gene, and hence they take different stances to respond to this situation (rf. Waters 2007). Some are gene skeptics (e.g., Kitcher 1992). Some take a dualistic position, such as Moss (2003), who distinguishes between Gene-P and Gene-D based on the fields in that gene concepts are applied. Some are pluralists, such as Griffiths and Stotz (2006, 2013), who differentiate between three senses of the gene: the instrumental gene, the nominal molecular gene, and the postgenomic molecular gene. Still others are both pluralists and pragmatists. Waters (2018) emphasizes that scientists do and should apply different gene concepts under various investigative contexts.

With some exceptions, few philosophers explore the reference problem of the term "gene". Although Fregean semantics holds that the sense/concept or intension of a name determines its reference or extension, the matter about how a sense determines the reference is not easily seen from the scientific context. The determination of a theoretical term's reference usually involves experimental procedures and techniques that should be investigated and analyzed. Weber (2005, ch.7) does impressive work by providing several reference-determining descriptions of the term "gene" in the history of genetics. Based on those descriptions and the analysis of *Drosophila* genetic practices, he suggests that the pattern of referential change for "gene" is a kind of

freely floating reference. He also argues that different gene concepts refer to *different* natural kinds, which are overlapping but not coextensive.[1] According to Weber, reference for "gene" is fixed in the following manner for classical and molecular genes.

> Reference of [classical] "gene" (2): Whatever (a) is located on a chromosome, (b) segregates according to Mendel's first law, (c) assorts independent of other genes according to Mendel's second law if these other genes are located on a different chromosome, (d) recombines by crossing-over, (e) complements alleles of other genes, and (f) undergoes mutations that cause phenotypic differences. (Weber 2005: 210)

> Reference of [molecular] "gene" (5): The class of DNA sequences that determine the linear sequence of amino acids in a protein. (Weber 2005: 212)

Both classical and molecular gene concepts do refer to natural objects, because, as Weber notes (2005: 210-211), some *tokens* satisfying the reference-determining descriptions are experimentally presented when using the concepts with the intention of referring to sets of entities in historical occasions. However, one should note that the experimented tokens in classical genetics seems to be only some organisms with specific phenotypes (say, fruit flies or other kinds of organisms) while the experimented tokens in molecular biology may be some DNA segments. This difference raises interesting problem: what tokens are individuated in different contexts of experiments?

Before moving to the next section, I want to clarify that the individuation problem of gene concept's tokens is not the issue of gene individuality as raised by Rosenberg (2006: 121-133).[2] He defends the gene individuality thesis in parallel to the species individuality thesis, but Reydon (2009) objects to his argument and defends the gene as a natural kind. This paper aims to discuss how a gene kind and its tokens are individuated rather than whether or not an allele such as *Hbf* (the human fetal hemoglobin gene) is an individual.

3. Chromosomal location of a gene

[1] Baetu (2011: 411) argues that "the referents of classical and molecular gene concepts are coextensive to a higher degree than admitted by Waters and Weber…" However, Baetu builds his argument in terms of Benzer's work on phage. In my view, he does not successfully refute Waters' and Weber's arguments, because the referential change occurred within the classical gene concept, as Weber cogently argues.

[2] Rosenberg uses "natural selection and the individuation of genes" as the title of the section in which he discusses the gene individuality thesis.

Weber's argument indicates that we may and should consider the reference of the classical gene concept independently of the molecular gene concept and others. Weber's reference-determining description of "gene" (2) indicates that the chromosomal location (or mapping) of genes plays a key role in determining referents. However, the question "what tokens are individuated and thus referred to?" does not be answered.

Classical geneticists in the early 20th century located and labeled some specific classical genes on some specific chromosomes. The earliest genetic map (see Figure 1) of *Drosophila melanogaster* (fruit fly) was depicted in 1915. Figure 1 shows that the gene (allele) pair of *Drosophila's* grey body and (mutant) yellow body is located at the first locus on the first chromosome. The second gene pair of red eyes and (mutant) white eyes is located below the grey body gene. The other genes are located below the first two in order. However, every gene is differently distant from the first gene and thus occupies *a single locus* without overlapping. Accordingly, are we able to say that the location of a gene individuates the gene? Before answering this question, it is necessary to discuss how classical geneticists locate a gene on a chromosome. In other words, what technique is used in the process of locating genes?

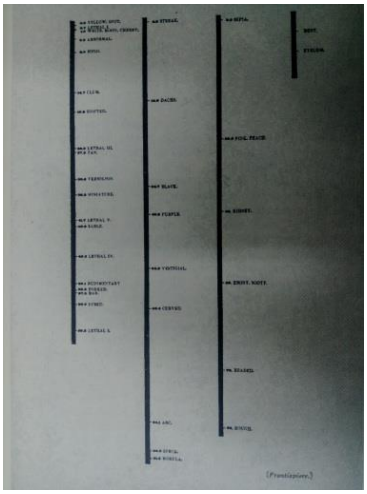


Fig. 1. Genetic map of *Drosophila* in 1915. Reproduced from Morgan, T. H. et. al. (1915).

Chromosomal location or mapping of genes is a well-known story (Darden 1991, Waters 2004, Weber 2005, 2006; Falk 2009). For the purpose of this paper, I introduce a very brief version. In the 1910s, Thomas Hunt Morgan's team developed a

Seattle, WA; 1-4 November 2018 -163-

technique to map the linear relations among factors (genes) in linkage groups, using Mendelian breeding data. Morgan and his team discovered that a pair of chromosomes may cross over with each other partially during the period of meiosis. Crossing over produces a specific ratio of the linked traits. Morgan believed that "the percentage of crossing over is an expression of the 'distance' of the factors from each other." (Morgan et.al. 1915: 61) Sturtevant then used percentages of linked characters that exhibited crossing over to calculate the relative positions of the factors to each other. This is the kernel technique for constructing genetic maps. By using genetic maps, Morgan's team determined the loci of many genes on the four chromosomes of *Drosophila*. Given the genetic maps, the classical geneticists assume that no other genes are located at the same position of a chromosome.[3] As a consequence, the single location of a gene actually indicates the individuality of genes.

Genetic maps by nature are diagrammatic models for the actual loci of genes in chromosomes. They are inferences from the statistical data of breeding experiments. Models represent the general. When we say that the location of a gene in a genetic map represents the locus of a classical gene on a chromosome, we really mean that it represents the locus of a type of classical gene on an identical type of chromosome in a cell within a kind of organism. Of course, this implies that a token of a type of classical gene on a token of a type of chromosome can be cognitively identified and discerned, because we can distinguish it from the tokens of the other genes. As a result, we can also count genes within cells. The located genes thus satisfy the two traditional characteristics of individuality: distinguishability and countability.[4]

If all chromosomes were stick-shaped substances of uniform material without complicated structure, then the chromosomal location of classical genes would be able to genuinely individuate them. According to molecular biology, however, chromosomes are a long chain of double helix DNA molecules that curl themselves up in twisted shapes. In such a case, we cannot delineate a located classical gene or depict its contour or boundary, because the chromosomal locus at which the gene is located includes a twisted part of the long DNA molecule. Even by invoking the knowledge from molecular biology, one would still be puzzled by the problem of defining the molecular gene.

4. Individuating molecular genes as individuals

Ever since the era of molecular biology, the continuously accumulating knowledge of genetics has not solved the individuation problem of genes. Instead, it

[3] Of course, a full story is more complicated. For the simplifying purpose, I skip the relevant discussion about gene mutation.

[4] The implications of using these criteria will be discussed in the sixth section.

has brought more troubles about the definition of the gene concept. Is a gene "a sequence of DNA for encoding and producing a polypeptide"? Should we include the start and stop codons (i. e., the regulation problem)? Should we count those introns deleted during the process of transcription into the investigated gene (i.e., the splicing problem)? The difficulty in defining the molecular gene concept directly contributes to the impediment of individuating a gene.

Many gene sequencing projects have been conducted during the genomic era. Scientists do not identify a DNA sequence as a gene and discern the gene from others by using gene sequencing *per se*, because it offers only syntactical orders of genetic codes. Gene annotation, which is used to infer what those annotated sequences do, has been developed to offer *senses* or *intensions* for them. However, the impediment of discerning genes remains, because the definition of the gene is still vague and confusing (rf. Baetu 2012; Gerstein et. al. 2007; Griffiths and Stotz 2013, ch. 4). In fact, gene annotation is based on several assumptions, by which scientists infer that a few sequences may be genes that contribute to phenotypes or functions. Those assumptions need to be confirmed by experimental investigations. Many techniques such as directed deletion, point mutation making, gene silencing, and transgenesis in reverse genetics have been developed to determine what a gene is and what it does (Gilchrist and Haughn 2010).

I argue that the transgenic technique is a very definite and powerful way to individuate a gene. It can even individuate molecular genes as individuals without a clear boundary of a gene or a clear definition of the gene, although the technique is limited.[5] How does the transgenic technique do this? What conditions of individuality allow the technique to individuate a gene as an individual?

Chen (2016) proposes a conception of experimental individuality with three attendant criteria (separability, manipulability, and maintainability of structural unity) and argues that the first experiment of bacteria transformation individuated an antibiotic resistance gene by satisfying the three criteria.[6] Below I reiterate this story in brief.

Stanley Cohen and Herbert Boyer combined DNA of *Escherichia coli* (*E. coli*) in 1973 and 1974 by transferring two different DNA segments encoding proteins for ampicillin and tetracycline resistance into *E. coli*, thereby realizing the transformation of this bacterium (Cohen et. al. 1973; Chang and Cohen 1974). Both DNA segments are called an "antibiotic resistance gene." Cohen and Boyer used small circular

[5] The technique cannot be applied in many occasions because of technological difficulties. It should not be applied to humankind due to ethics consideration. In addition, many gene-modification organisms produced by using the technique may involve ethical issues.

[6] Chen (2016) uses the creation of Bose-Einstein condensates in physical experiments as the other example. Chen's intent is to argue that biological entities and physical entities in laboratories share the same criteria of experimental individuality.

Seattle, WA; 1-4 November 2018 -165-

plasmids (extrachromosomal pieces of DNA) as vectors to transfer a foreign DNA segment into a bacterial cell. The plasmids were made by cutting out a (supposed) antibiotic resistance gene from other bacteria with the restriction enzyme *EcoRI*, linking the segment into a plasmid by using another enzyme, DNA ligase. The scientists then transferred the plasmid into an *E. coli* cell without the ability to resist antibiotics. The result, a modified *E. coli* cell, was able to resist antibiotics and contained the antibiotic resistance gene. In that experiment, the antibiotic gene was separated from its original bacteria and then was manipulated (i.e., linked and transferred). Its structural unity was not broken down, hence allowing it to be expressed in the other kind of bacteria. Scientists thus identify it as a gene, an individual biological entity, because the separated, manipulated, and maintained antibiotic gene was naturally separable, manipulable, and maintainable. The photos in Figure 2 show that scientists worked with a single DNA segment, as indicated by (b) in [A] and [B].

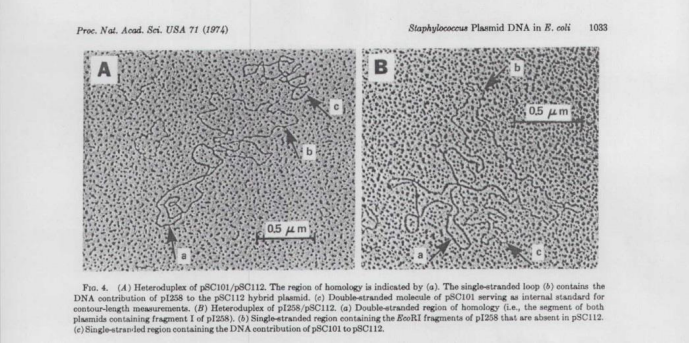


Fig. 2. Two pictures of plasmids in bacterial transformation. Reproduced from Chang and Cohen (1974).

I next interpret the performance of the technique used in transgenic experiments as the general process of individuating transgenes. The process has five stages.

(1) Use restriction enzymes to cleave specific segments from recognition sites of long DNA chains. A specific restriction enzyme can cut away a specific DNA segment at a specific site.

(2) Link the cleaved segment of DNA to a plasmid vector by using DNA ligase. The vector is a circular DNA that may come from a wild type of virus.

(3) Incorporate the DNA segment in the vector into the genome of another organism by injecting the plasmid vector to a cell of the target organism. Of course,

they may fail when the intended feature is not expressed.

(4) Make copies of DNA segments by cloning the cell containing the transferred segment of DNA. The aim of DNA cloning is to copy a segment of interest (or a gene) from an organism and produce many copies.

(5) Observe the expression of the novel feature that the target organism does not typically have. If a DNA segment cut from an original organism is successfully pasted into a cell of a target organism and the target organism expresses the intended feature that the original organism has, then one concludes that the segment is a gene.

The first stage corresponds to the separation condition, the second, the third, and the fourth stages to the manipulation condition, and the fifth stage to the maintenance condition. Accordingly, one can easily see that those cut, linked, transferred, pasted, and copied genes are particulars – individuals, because they satisfy the three criteria of experimental individuality that indicates their *singularity* and *particularity*. In other words, a single segment of DNA maintains its structural unity when being separated and manipulated. This is so, because cutting a gene from an original organism is in fact separating it from its environment and because transferring, pasting and copying a gene is manipulating it. If the gene does express the intended feature in a target organism, then this condition indicates that the unity of its chemical and informational structure has been maintained.

5. Two kinds of individuation of genes

The previous discussion indicates that two different objects have been individuated in different experimental and theoretical contexts. In the context of classical genetics, scientists used breeding experiments and theoretical inferences to locate a gene at some locus on a chromosome. They would individuate genes as types if they assume that no other genes could coexist at the same locus. If one interprets the meaning of "individuation" as "only individuals can be individuated," then the phrase "individuating genes as types" sounds unreasonable. Is it better to say "unitization of genes" rather than "individuation of genes"?

It is quite right to say classical geneticists *unitize* genes as types. In a sense, however, we may reasonably say that we individuate a gene as a type, because the type has tokens or members that are distinguishable and countable individuals. Classical geneticists suppose that all types of genes have corpuscular members, i.e., substantive individuals. In such a sense, talking of "individuating genes as types" is reasonable. If no distinguishable and countable members or samples of a kind can be identified, then the kind cannot be individuated. In other words, we cannot individuate

Seattle, WA; 1-4 November 2018 -167-

such a kind as water or air that is expressed by "mass" nouns at the macroscopic level, although we can individuate a sample of water by using a container or individuate a water molecule by specific technique at the molecular level. For the cases of experiments using the transgenic technique, molecular biologists physically individuate *singular and particular* gene tokens. Thus, we claim that scientists experimentally individuate genes as individuals in such a context.

In consequence, two different sets of criteria for individuality are presupposed. Experiments using the location technique have individuated a type whose tokens or members are countable individuals rather than matter referred to by mass nouns. In such experimental contexts, we emphasize distinguishability and countability as the indexing features of individuals. Experiments using the transgenic technique individuate singular and particular individuals – gene tokens. For these experimental contexts, we emphasize singularity and particularity of individuals in contrast to universality of types or kinds. We assure the particularity and singularity of the individuals through the realization of experimental individuality, namely, the joint realization of separability, manipulability, and maintainability of structural unity. At this point, more philosophical implications will be discussed in next section.

The two individuated targets indicate two different referential levels of the term "gene" in the literature. As we have seen, when many philosophers and scientists ask "what is a gene," they really refer to a type of gene in conjunction with discussing the gene concept or the definition of "gene." Similarly, in some contexts of scientific investigation, scientists use "a gene" to refer to a type of gene as the phrase "chromosomal location of a gene". In the context of transgenic experiments, however, "a gene" is used to refer to a genuine individual – a single and particular gene token, because scientists have worked with particular objects that maintain their structural unity when being separated and manipulated in the process of experimenting.

The two referential levels indicate two different kinds or levels of experimental individuation, which are realized by two different techniques: the chromosomal location technique and the transgenic technique. Although the two techniques aim to the same target (i.e., genes or types of genes), they physically experiment and manipulate different objects. Experiments using the chromosomal location technique indirectly identify loci of genes by manipulating organisms that contain chromosomes with genes in breeding, while experiments using the transgenic technique directly manipulate DNA segments. Therefore, classical geneticists can only cognitively discern gene types by identifying their loci without practically interacting with gene tokens; they really practically interact with organismal individuals that contain different types of genes. Reversely, molecular biologists can practically interact with gene tokens and then cognitively infer out the existence of a gene type.

6. Gene concepts and individuation

One may still wonder: Can the location technique individuate a singular and particular gene in the sense of individuating entities as individuals? The answer is obviously negative, because that technique cannot separate and manipulate a gene token and maintain its structural unity. On the contrary, one may ask: Can the transgenic technique individuate a type of gene? Here the answer is less clear. In the sense that scientists suppose that a token of a gene has been physically individuated in transgenic experiments, we are allowed to say that the technique also individuates a type of gene. However, scientists are not fully sure that the transgenic technique on a posited gene can be always successfully applied to another individual of the same organism. In fact, the probability of failure is quite high. Unless the experimental individuation of particular tokens can be performed repeatedly and stably, then one can say that the gene tokens indicate a general type of gene and that the type has been identified. However, the object individuated by the technique is not a type of gene, because the technique always requires manipulating particular segments of DNA -- gene tokens. If a kind of transgenic experiment with a specific transgene has been stably repeated, then a type of gene has been discovered by experimentally individuating its tokens in performing such an experiment.

Since transgenic experiments may be successfully and stably performed by using different transgenes, one can extract a special conception of the gene that is characterized by the transgenic technique. I call this "the transgenic conception of the gene," in which *a gene is a transferrable DNA sequence which is able to express a phenotype/function on another kind of organisms*. Of course, this does not imply that those technically untransferrable DNA sequences are not genes, given the fact that the number of transgenes is relatively few to the number of genes located at chromosomes. This is so because scientists do not always find the precise site of a gene (type) and available restriction enzymes to cut the DNA segment of the gene. Thus, the extension of the transgenic conception of the gene is not equivalent to that of the classical gene concept. Due to the limited number of transgenes, the transgenic conception is not yet co-extensional with the molecular gene concept. To be precise, the extension of the former is included within the extension of the latter, because all transgenes are molecular genes but not all molecular genes can be transplanted. In addition, the intension of the transgenic conception is implied in the intension of the molecular gene concept, because the technique was developed from molecular biology. As a consequence, the transgenic conception can be viewed as a *sub-conception* of the molecular gene concept. Nevertheless, we have a conception

Seattle, WA; 1-4 November 2018 -169-

derived from scientific practices.

7. The priority of individuation to individuality

Bueno, Chen, and Fagan (2018) promote an approach by which investigating processes of individuation in scientific practices is prior to metaphysical speculation on criteria of individuality. This paper obviously follows the approach. However, this does not mean that we do not need any criterion of individuality in identifying any individual in scientific practices. Rather, criteria of individuality are implied in or extracted from procedures of scientific practices, as the three conditions of experimental individuality are extracted from experimental practices (Chen 2016). Criteria of individuality based on scientific practices may or may not conflict with criteria from metaphysical theories. Considering the relationship between practical criteria and speculative criteria will help us understand practical individuation more deeply.

The metaphysical tradition has identified at least six characteristics or indexing features of individuality in general: particularity, distinguishability, countability, delineability, unity, and persistence (Pradeu 2012: 228-229; Chen 2016: 351).[7] Recently, some philosophers argues that all biological entities are processes (Dupré 2018, Nicholson and Dupré 2018, Pemberton 2018), so I would like to add processuality to the list. Indeed, I believe that all biological individuals pass through a life, i.e., a process (see also Chen 2018), therefore, processuality is a central characteristic of biological individuality. Those characteristics, originally come from metaphysical speculation, can singly, jointly, or collectively serve as epistemic criteria of individuality.

In the context of scientific practices, they are the outcomes from rather than preconditions for the realization of individuation. For example, individuating genes as individuals in the context of transgenic experiments indicates that the separated, manipulated, and maintained genes are particular and singular tokens. As the experimental individuation of gene tokens is realized, those tokens are also distinguishable, countable, unitary, persistent, and passing through a process, because particular and concrete individuals are being separated, manipulated, and maintained. The practices of separation and manipulation indicate epistemic particularity,

[7] Characteristics of individuality can serve as criteria of individuality and thus be involved in a theory of individuation. Bueno, Chen, and Fagan (2018) identify six theories of individuation in traditionally analytic metaphysics. A theory of individuation in the metaphysical sense involves not only "a theoretic construction of the nature of individuality and its attendant criteria," but also other metaphysical concepts such as "property, trope, universal, particular, substance, substratum, time, space, sort or kind." (p. 3) For my purpose, I will discuss only characteristics of individuality rather than any theory of individuation.

distinguishability, and countability. The practice of maintenance of structural unity indicates the unity, persistence, and processuality of the maintained gene token. However, all of the three practices would not indicate the delineation of a gene token, because the spatial boundary of the manipulated gene does not and cannot be delineated. Of course, this point does not mean that delineation is not a characteristic of individuality, but rather that it is not applicable to this case.

Individuating genes as types in classical genetics indicates that the individuated types of genes contain distinguishable and countable tokens, because the individuation is the location of a gene at a chromosome in a diagrammatic model. Supposing that the loci of different genes do not overlap, then the special locus of a gene is thus distinguishable from the locus of another gene. As a consequence, a gene token at a chromosome in a cell of a kind of organism is thus distinguishable from another token of the identical type of gene. All gene types located at chromosomes are countable. Supposing that every organism contains a token of a specific type of gene, then tokens of that gene type are countable. However, chromosomal location of genes does not indicate particular and singular gene tokens, because the individuated objects are only types of genes. As I have argued, the kind of individuation practice did not touch down the manipulation of individuals and remained in the cognitive level which focuses on gene types in general.

Although the concept of individuation can be reasonably applied to a kind whose members are individuals, all characteristics of individuality are not applicable. One cannot apply particularity, delineation, unity, and processuality to gene types, because a gene type is, in principle, universal, occupying multiple spaces, not cohesive, replicable, and non-processual. However, distinguishability and countability can be adequately applied to gene types, because one can distinguish one gene type from another gene type and count gene types when the chromosomal location is realized. In this case, thus, both distinguishability and countability cannot sufficiently demonstrate that the individuated objects are individuals. On the other hand, in the case of transgenic experiments, we can derive particularity, unity, and processuality from the three conditions of experimental individuation (separation, manipulation, and maintenance of structural unity). As a consequence, characteristics of individuality are derived from individuation; they are outcomes of practical individuation.

8. Conclusion

In this paper, I argue that there are at least two kinds of experimental individuation of genes. Scientists individuate genes as types in classical genetics and

individuate genes as tokens in transgenic experiments. Individuating a gene as a type or individuating a gene as an individual depends on the technique used in experimentation. I argue that characteristics of individuality identified in traditional metaphysics are not presupposed by individuation. Rather, they are outcomes or products derived from practical individuation in scientific experiments. I further argue that different kinds of experimental individuation presuppose different concepts of the gene: the classical gene concept and the transgenic conception of the gene. I argue that the transgenic conception can be viewed as a sub-conception of the molecular gene concept. An outstanding problem remains. Whether we can unify different concepts of the gene by integrating different experimental techniques, such as the chromosomal location technique, the technique of genetic sequencing, the techniques in reverse genetics, and the transgenic technique. Future analyses can approach this and other related questions in light of our new understanding of how classical geneticists individuated genes and the role experimental techniques play in identifying a gene as an individual.

Acknowledgment: I thank Alan Love, Ken Water, and Marcel Weber for their very valuable comments and suggestions. This paper has been revised according to their comments.

References

Baetu, Tudor M., 2011. “The referential convergence of gene concepts based on classical and molecular analysis,” *International Studies in the Philosophy of Science*, 24 (4): 411-427.

Baetu, Tudor M., 2012. “Genes after the human genome project.” *Studies in History and Philosophy of Biological and Biomedical Science*, 43: 191-201.

Beurton, P., R. Falk, and H.- J. Rheinberger, 2000. *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*. Cambridge, UK: Cambridge University Press.

Beuno, Otavio, Ruey-Lin Chen, and Melinda B. Fagan, 2018. “Individuation, process, and scientific practice.” In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 1-18. New York: Oxford University Press.

Carlson, E., 1991. “Defining the gene: an evolving concept.” *American Journal of Human Genetics*, 49: 475-487.

Chang, Annie C. Y. and Stanley N. Cohen, 1974. “Genome construction between bacterial species *in vrito*: Replication and expression of *Staphylococcus* plasmids

in *Escherichia coli*," *Proceedings of the National Academy of Science of USA*, 71(4): 1030-1034.

Chen, Ruey-Lin, 2016. "The experimental realization of individuality." In Alexandre Guay and Thomas Pradeu (eds.). *Individuals across the Sciences*, 348-370. New York: Oxford University Press.

Chen, Ruey-Lin, 2018. "Experimental Individuation: Creation and Presentation," In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, . New York: Oxford University Press.

Cohen, Stanley N. et. al., 1973. "Construction of biologically functional bacterial plasmids *in vitro*," *Proceedings of the National Academy of Science of USA*, 70(11): 3240-3244.

Darden, Lindley, 1991. *Theory Chang in Science: Strategies from Mendelian Genetics*. Oxford: Oxford University Press.

Dupré, John, 2018. "Processes, Organisms, Kinds, and Inevitability of Pluralism." In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 25-38. New York: Oxford University Press.

Falk, Raphael, 1986. "What is a gene?" *Studies in History and Philosophy of Science*, 17: 133-173.

Falk, Raphael, 2009. *Genetic Analysis: A History of Genetic Thinking*. Cambridge: Cambridge University Press.

Falk, Raphael, 2010. "What is a gene – revised" *Studies in History and Philosophy of Biological and Biomedical Science*, 41: 396-406.

Gerstein, Mark B. et. al. 2007. "What is a gene, post-ENCODE? History and updated definition." *Genome Research*, 17(6): 669-681.

Gilchrist, Erin and George Haughn, 2010. "Reverse genetics techniques: engineering loss and gain of gene function in plants," *Briefings in Functional Genomes*, 9(2): 103-110.

Griffths, Paul and Karola Stotz, 2006. "Genes in the postgenomic era," *Theoretical Medicine and Bioethics*, 27(6): 253-258.

Griffths, Paul and Karola Stotz, 2013. *Genetics and Philosophy: An Introduction*. Cambridge: Cambridge University Press.

Kitcher, P. S., 1982. "Genes." *British Journal for the Philosophy of Science*, 33: 337-359.

Kitcher, P. S., 1992. "Gene: current usages." In E. Keller and L Lloyd (eds.), *Keywords in Evolutionary Biology*. Cambridge, MA: Harvard University Press, pp. 128-131.

Lowe, E. Jonathan 2005. "Individuation," *The Oxford Handbook of Metaphysics*, ed. Michael J. Loux and Dean W. Zimmerman. Oxford: Oxford University Press.

Maienchein, J., 1992. “Gene: Historical perspectives.” In E. Keller and E. Lloyd (eds.). *Keywords in evolutionary biology*. Cambridge, MA: Harvard University Press, pp. 181-187.

Morgan, Thomas Hunt, et.al., 1915. *The Mechanism of Mendelian Heredity*. New York: Henry Holt and Company.

Moss, Lenny, 2003. *What Genes Can’t Do*. Cambridge, Mass.: The MIT Press.

Nicholson, Daniel J. and John Dupré, 2018. Everything flows: Towards Processual Philosophy of Biology.

Pearson, Helen, 2006. “What is a gene?” *Nature*, 441(25): 399-401.

Pemberton, John. 2018. “Individuating Processes,” In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 39-62. New York: Oxford University Press.

Pradeu, Thomas, 2012. *The Limits of the Self: Immunology and Biological Identity*. Oxford: Oxford University Press.

Reydon, Thomas, 2009. “Gene Names as Proper Names of Individuals: An Assessment.” *British Journal for the Philosophy of Science*, 60(2): 409-432.

Rosenberg, Alexander, 2006. *Darwinian Reductionism*. Chicago: The University of Chicago Press.

Snyder, Michael and Mark Gerstein, 2003. “Defining genes in the genomics era.” *Science*, 300(5617): 258-260.

Stotz, Karola and Paul Griffths, 2004. “Genes: philosophical analyses put to the test.” *History and Philosophy of the Life Sciences*, 26: 5-28.

Wain, H. M., et. al. 2002. “Guidelines for human genome nomenclature,” *Genomics*, 79: 464-470.

Waters, Kenneth C., 1994. “Genes made molecular,” *Philosophy of Science*, 61: 163–185.

Waters, Kenneth C., 2004. “What was classical genetics?” *Studies in History and Philosophy of Science*, 35 (4): 783-809.

Waters, Kenneth C., 2007. “Molecular genetics,” *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/molecular-genetics/

Waters, Kenneth C., 2018. “Don’t Ask ‘What is an individual?’” In Otavio Beuno, Ruey-Lin Chen, and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 91-113. New York: Oxford University Press. (In press)

Weber, Marcel, 2005. *Philosophy of Experimental Biology*. Cambridge, UK: Cambridge University Press.

Weber, Marcel, 2006. “Representing genes: Classical mapping techniques and the growth of genetic knowledge,” *Studies in History and Philosophy of Biological and Biomedical Science*, 29: 295-315.

The Verdict's Out:

Against the Internal View of the Gauge/Gravity Duality

4993 words

Abstract

The gauge/gravity duality and its relation to the possible emergence (in some sense) of gravity from quantum physics has been much discussed. Recently, however, Sebastian De Haro (2017) has argued that the very notion of a duality precludes emergence, given what he calls the internal view of dualities, on which the dual theories are physically equivalent. However, I argue that De Haro's argument for the internal view is not convincing, and we do not have good reasons to adopt it. In turn, I propose we adopt the external view, on which dual theories are not physically equivalent, instead.

1 Introduction

The gauge/gravity duality has generated much discussion about whether space-time geometry or gravity emerges (in some sense) from quantum physics.[1] Recently, however, De Haro [2017] has argued that the very notion of a duality *precludes* the possibility of emergence given what he calls the *internal view* of dualities, on which dual theories are physically equivalent. In turn, this claim impinges upon the broader debate about whether we can make claims about emergence given a duality. After all, since the internal view of dualities is supposed to *rule out* emergence, any such debate is rendered moot once we adopt the internal view. My goal here, though, is to argue that De Haro's argument for the internal view is not convincing. Instead, I propose we adopt the *external view* of dualities, on which dual theories are *not* physically equivalent.

First, I introduce Fraser's [2017] three-pronged distinction of predictive, formal and physical equivalences, characterizing dualities in terms of this distinction (§2.1). I then make things more concrete by briefly considering the gauge/gravity duality via the Ryu-Takayanagi conjecture from the **AdS/CFT** (anti-de Sitter space/conformal field theory) correspondence (§2.2).

Next, I introduce De Haro's interpretive fork between the internal and external views of dualities (§3). I illustrate how the internal view is supposed to preclude emergence, but criticize De Haro's argument for the internal view – that it is meaningless to hold the external view given 'some form of' structural realism and how the two theories are

[1] One prominent physicist who is a proponent of emergent space-time is Seiberg 2007, while philosophers like Rickles 2011/2017, Teh 2013, and Crowther 2014 have all tackled the topic.

'totalizing' in some way – by showing how it does not work without further assumptions (§4). In turn, given the interpretive fork, I propose we adopt the external view instead. In concluding remarks, I briefly discuss this result in relation to the broader debate about emergence within the gauge/gravity duality.

2 Gauge/Gravity through AdS/CFT

2.1 Duality

Fraser [2017] takes two theories related by a duality to have two features: (i) they agree on the transition amplitudes and mass spectra, and (ii) there is a 'translation manual' that allows us to transform a description given by one theory to a description given by another theory. We may explicate (i) and (ii) by first considering distinct sorts of 'equivalence' proposed by Fraser [2017, 35]:

- *Predictive equivalence*: "there is a map from T_1 to T_2 that preserves the values of all expectation values deemed to have empirical significance by T_1 and that preserves the mass spectra, and vice versa."
- *Formal equivalence*: "there is a translation manual from T_1 to T_2 which maps all quantum states and quantum observables deemed to have physical significance by T_1 into quantities in T_2 and respects predictive equivalence, and vice versa."
- *Physical equivalence*: "there is a map from T_1 to T_2 that maps each physically significant quantity in T_1 to a quantity in T_2 with the same physical interpretation and respects both formal and predictive equivalence, and vice versa."

Seattle, WA; 1-4 November 2018 -177-

Given our characterization of a duality as (i) and (ii), we may quite naturally say that two theories are dual to one another when they are *predictively* and *formally* equivalent. Furthermore, supposing that this three-pronged distinction exhausts the possible equivalences relevant to physics, we might also say that two theories satisfying (i)-(iii) are also *fully*, or *theoretically*, equivalent.

Here it would be germane to differentiate two distinct sorts of structures in a duality. Given predictive and formal equivalence, the isomorphism holding between physical and empirical quantities of the dual theories suggests a structure, which may be called the *empirical core* of the duality. However, as Teh [2013, 301] also notes, despite the empirical core, "duality is precisely an equivalence between two theories that describe (in general) different physical structures, i.e. theories with non-isomorphic models." In other words, while there is an empirical core, by which physical and empirical quantities are mapped onto one another, these quantities are generally related to other quantities in a quite different manner on each side, viz. there is 'excess structure' exogenous to the empirical core. Without further argument, we are not entitled to 'discard' this 'excess structure', which also means that predictive and formal equivalence (characterizing the empirical core) does not automatically entail physical, and hence full, equivalence.

Given Fraser's framework, I will briefly introduce the gauge/gravity duality more concrete by briefly examining the example of **AdS/CFT** correspondence.

2.2 The AdS/CFT Correspondence

The *gauge/gravity duality*, or *holographic principle*, postulates a duality between a suitably chosen N-dimensional gauge quantum field theory (QFT) that does not describe

gravity, and a quantum theory of gravity in $(N+1)$-dimensional space-time (the 'bulk') with an N-dimensional 'boundary', on which the gauge theory is defined. Hence the slogan: gauge on the boundary, gravity in the bulk.

The **AdS/CFT** correspondence is a specific case of the gauge/gravity duality. On the one hand, '**AdS**' stands for anti-de Sitter space-time - a maximally symmetric solution to the Einstein equations with a constant negative curvature and a negative cosmological constant. More accurately, though, the '**AdS**' in **AdS/CFT** correspondence should be taken to refer to a string theory of quantum gravity defined *on* a 5-dimensional **AdS**. '**CFT**', on the other hand, refers to a quantum field theory with scale (or conformal) invariance defined on the 4-dimensional boundary of the **AdS**. The **AdS**-side theory is defined in the 'bulk', and the **CFT**-side theory is defined on the 'boundary' of the **AdS** space-time.

The **AdS/CFT** correspondence, then, refers to a postulated duality between the two theories, satisfying (i) and (ii) from §2.1. (i) is satisfied given the postulate that bulk fields propagating in the bulk are coupled to operators in the boundary **CFT**. Hence, the **AdS** theory of gravity will predict exactly the 'same physics', viz. transition amplitudes, expectation values and so on, as the **CFT** theory without gravity.

Beyond empirical, i.e. measurable, quantities, physically significant quantities of **AdS/CFT** must also relate to one another since it is a duality. In other words, (ii) is supposed to hold simply as a core postulate. This is not to say that (ii) is completely unfounded: in particular, we have evidence suggesting that at least *some* physical quantities of dual theories are related to one another in surprising ways, which in turn supports the claim that (ii) holds. Here I will focus on one such relation, the Ryu-Takayanagi conjecture.

The Ryu-Takayanagi conjecture postulates that the entanglement entropy of two regions on the boundary is related to the surface area within the bulk:[2]

$$(\mathbf{RT}):\ S_A = \frac{Area(\widetilde{A})}{4G_N}$$

RT tells us that the entanglement entropy of a region on the boundary of the **AdS**, S_A, viz. the von Neumann entropy[3] in the **CFT**, is directly proportionate (by 4 times the Newtonian gravitational constant) to the area of the boundary surface $\widetilde{A}$ bisecting the bulk, dividing the two entangled regions on the boundary. Below, *Fig. 1.* shows a simplified diagram for visualizing **RT**.

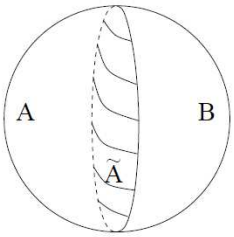


Fig. 1. The area $\widetilde{A}$ bisects the bulk space-time into two, and on the boundaries of the two parts we define the regions A and B. The Ryu-Takayanagi formula tells us that given a change in S_A we get a change in the size of $\widetilde{A}$ by the proportion of $\frac{1}{4G_N}$. [Figure taken from Van Raamsdonk 2010]

RT paints an interesting picture for emergence of space-time geometry from quantum theory: the area of a space-time itself is closely related to quantum entanglement entropy in a surprising way. An increase in the entanglement entropy between two

[2]See Ryu & Takayanagi 2006 for technical details.

[3]The von Neumann entropy is given by $S_A = -Tr(\rho_A log \rho_A)$. The reduced density matrix describing the region A, ρ_A, is obtained from tracing over the B-components of the combined density matrix of A and the entangled region B, ρ_{AB}: $\rho_A = Tr_B(\rho_{AB})$.

regions of a field described by **CFT** leads to a proportionately increasing boundary area of the bulk, and hence a geometric (or gravitational) phenomenon is described in terms of a quantum phenomenon.[4]

Given relations like **RT**, we can also see more clearly how **AdS/CFT** is supposed to satisfy (ii): physically significant quantities, such as 'area' of space-time in the bulk and 'entanglement entropy' between two regions on the boundary, are mapped to one another via suitable equations. Hence, **AdS/CFT** is a special case of the gauge/gravity duality: a theory of quantum gravity on a $(N+1)$-dimensional **AdS** space-time is dual to a **CFT** defined on its N-dimensional boundary.

With the gauge/gravity duality made concrete, let us turn to the interpretive task.

3 The Internal View

Dieks et al. [2015] and De Haro [2017] proposes an interpretive fork for dualities: we can either adopt an internal or external view. De Haro describes the *internal view* as such:

> if the meaning of the symbols is not fixed beforehand, then the two theories, related by the duality, can describe the same physical quantities. [...] we have two formulations of one theory, not two theories. [De Haro 2017, 116]

On the contrary, the *external view* holds that:

> the interpretative apparatus for the entire theory is fixed on each side. [...] On this interpretation there is only a formal/theoretical, but no empirical, equivalence between the two theories, as they clearly use different physical

[4]See Van Raamsdonk 2010 for an excellent summary of this picture.

> quantities; only one of them can adequately describe the relevant empirical observations.

Is De Haro's characterization of the external view adequate? The fact that there is no 'empirical' equivalence (what Fraser calls physical equivalence) between two theories does not entail that at most one of them can adequately describe the relevant empirical observations, where one description is 'correct' and the other 'wrong', nor does it entail mutually exclusive physics where only one theory can be correct at any one time. To assume so seems to rule out, by fiat, the possibility of emergence, since emergence relies on *both* theories being in a way adequately descriptive of the world (except one is more 'fundamental' than the other). Hence, taking in account Fraser's framework, I re-characterize the external view as such: it is simply the claim that the two dual theories are *physically non-equivalent* i.e. have distinct physical interpretations, despite formal and predictive equivalence.

Given the interpretive fork, if we are led to forsake the internal view, then we are motivated to accept the external view instead. As such, my strategy here is to show that we should forsake the internal view, and in turn accept the external view instead.

To better understand what the internal view is claiming, I break it down into three constituent claims.

The first claim is that of *theoretical equivalence*: under the gauge/gravity duality, the two theories (e.g. **AdS** and **CFT**) are taken to be simply different formulations of a single theory, describing the same physical quantities despite their obvious differences. As Dieks et al. puts it, 'the two theories collapse into one' [2015, 209-210]. In light of Fraser's framework described in §2.3, this claim means that the gauge/gravity duality, on

the internal view, involves the conjunction of predictive, formal and physical equivalences. In other words, beyond a one-to-one mapping (a 'translation manual') of relevant physical quantities and the sharing of all transition amplitudes, mass spectra and other observable predictions, the internal view claims that the two theories also have the *same physical interpretation*. However, as Fraser [2017, 35] notes, "predictive equivalence does not entail formal equivalence, and formal equivalence does not entail physical equivalence." Formal and predictive equivalence cannot entail physical equivalence on their own.

The internal view's claim of theoretical equivalence, then, must require an additional claim of *physical equivalence*, in addition to formal and predictive equivalence: the dual theories are taken to be physically equivalent, and hence have the same physical interpretation. As per §2.1, this would indeed entail theoretical equivalence.

Physical equivalence is in turn justified by a third claim, that the two theories in a duality should be left *uninterpreted*. As De Haro claims above, assume 'the meaning of the symbols is not fixed beforehand'. Then, given formal and predictive equivalence, we have an isomorphism between the dual theories' (now-uninterpreted) 'physical quantities' and numerical predictions, viz. an uninterpreted empirical core. Ignoring the 'excess structure' exogenous to the empirical core, we can then take the empirical core to be representing a single uninterpreted theory, where the now-uninterpreted 'quantities' of each dual theory now refer to the 'places' or 'nodes' of the empirical core's structure. As Dieks et al. (2015) puts it,

> A in one theory will denote exactly the same physical quantity as B [...] if these quantities occupy structurally identical nodes in their respective webs

Seattle, WA; 1-4 November 2018 -183-

> of observables and assume the same (expectation) values. [Dieks et al. 2015, 209]

Now, given this situation, it might seem plausible to claim that the dual theories are really physically equivalent. Consider **RT**. On the internal view, we are led to say that 'area' really has the same meaning as 'entanglement entropy'. After all, in the theoretical structure that is supposed to matter on the internal view, viz. the empirical core, the two terms are related structurally in the same way to other terms elsewhere (sans a proportional constant). Given that the two theories is also stripped of all prior physical meaning, this structural identity suggests that the 'area' and 'entanglement entropy' are really describing the same quantities, despite their obvious non-isomorphism more generally (e.g. different equations in computing these quantities in their respective theories, the terms involved in calculating them, and so on). In other words, it seems that we are allowed to proclaim physical equivalence on this view.

If we do accept this third claim, we get physical and hence theoretical equivalence, and so the internal view does preclude the possibility of emergence: Theoretical equivalence effectively rules out any account of emergence. If the two dual theories are really just different formulations of one theory, then there is nothing for this new, unified, theory to emerge *from*: nothing can emerge from itself in any interesting way. Subsequently, a duality is supposed to *preclude* emergence on the internal view.

Agreed: physical equivalence entails theoretical equivalence, and theoretical equivalence rules out any sort of emergence. However, are we forced to adopt physical equivalence given the internal view? De Haro himself seems unclear on this point. Note the use of "can" in his characterization of the internal view above: "the two theories,

related by the duality, *can* describe the same physical quantities" [2017, 116, emphasis mine]. Are we supposed to believe that physical equivalence *can* hold, or that it *must* hold, on the internal view? In other words, since physical equivalence hangs on the third claim of leaving terms of the dual theories uninterpreted, *must* we adopt the third claim, or is it merely *possible*?

De Haro seems to suggest that theoretical, and hence physical, equivalence *must* hold, since he assumes the two dual theories to be 'two formulations of *one theory*' [emphasis mine]. However, later on, he suggests that physical equivalence merely *can* hold, when he considers an example of leaving dual theories uninterpreted beyond structural relations:

> For what might intuitively be interpreted as a 'length, a reinterpretation in terms of 'renormalisation group scale is now *available*.[5] [De Haro 2017, 116, emphasis mine]

The *availability* of an interpretative stance – in our case of **RT**, of interpreting bulk boundary surface area to be the same physical quantity as entanglement entropy – surely does not entail the *necessity* of the stance. Hence, there are two readings of the internal view: on the weak reading, we take the modal talk – e.g. a reinterpretation being 'available' or how we 'can' describe the same physical quantities – seriously, and on the strong reading we ignore the modal talk completely.

On the one hand, the claim that the internal view precludes emergence is not true on the weaker view. On this view, *if* we assume that the terms on both sides of the duality are uninterpreted, then there is no emergence; *but* this is not forced on us. In turn, this

[5]For context, though unmentioned in this paper, length and renormalisation group scale are also dual quantities in AdS/CFT.

makes the preclusion of emergence merely possible. However, this reading of the internal view does not rule out emergence as De Haro claims. I will thus assume that De Haro intends for us to take the strong reading of the internal view, which does claim that the terms of the both sides *are* uninterpreted.

However, we have not yet seen a compelling reason for accepting the claim that we *have to* see the terms of the dual theories as uninterpreted, and subsequently that physical equivalence *must* hold. *A fortiori* we are not obliged to accept the internal view.

Indeed, something is odd about the argument structure I mapped out: To establish the second claim of physical equivalence, we must establish the third claim, that we must discard anything beyond the empirical core and to leave the terms uninterpreted. However, to justify leaving the terms uninterpreted requires a convincing argument for assuming physical equivalence between the two theories to begin with! Otherwise, we have no reason to simply discard the 'excess' structure and leave the dual theories' terms uninterpreted.

Hence, further arguments are required to establish the third claim. Furthermore, if we discover that this argument is wanting, we shall then have reasons to reject the internal view.

4 De Haro's Argument

De Haro does provide an argument, which runs on the idea that two plausible commitments entails the internal view: the commitment that the dual theories are theories of the whole world in some suitably totalizing manner, and the commitment to "some form of structural realism" [2017, 116].

Let us begin by examining the two commitments. The first commitment implies that dual theories are theories of the whole world, in the sense that they are "both candidate descriptions of the same world" [Dieks et al. 2015, 14]. However, *prima facie* this is not true, since on one hand we have a theory of gravity/space-time geometry, while on the other we have a theory without (not to mention different dimensionalities). How can two theories, one describing something the other does not, both be about the same world? We can try to make this assumption intelligible by taking into account the translation manual between the two theories. Given the translation manual, we can claim that the **CFT** theory without gravity does describe gravity in a way. Consider **RT**: while the entanglement entropy described within **CFT** does not appear to describe space-time geometry *by itself*, the **CFT** plus the translation manual *and* **AdS** (in this case **RT**) *does* describe space-time geometry, albeit in a higher-dimensional space-time. When the entanglement described within the **CFT** changes, the boundary surface area in the **AdS**-side theory with gravity changes as well. Hence, by considering the translation manual given by the duality, the first commitment is made plausible.

The second commitment requires us to adopt some form of structural realism. Structural realism here can be understood loosely, since nothing turns on the particular account of structural realism we employ. Furthermore, De Haro himself does not specify precisely what he means by 'some form of' structural realism. As such, I will likewise adopt a loose notion of structural realism: I understand it to be the view that we should be (metaphysically or epistemically) committed only to the mathematical or formal structure of our theories, and this entails, among other things, that theoretical terms are to be defined in terms of their relations to other places or nodes in this formal structure.

Now, De Haro then claims that the two commitments entail the internal view:

Seattle, WA; 1-4 November 2018 -187-

> If [the two commitments] are met, it is impossible, in fact meaningless, to decide that one formulation of the theory is superior, since both theories are equally successful by all epistemic criteria one should apply. [De Haro 2017, 116]

Since he does not flesh out his argument in much detail, I attempt to reconstruct his argument in a plausible fashion: firstly, let us grant the two commitments. Do these commitments commit us to the conclusion that it is meaningless to differentiate between the two dual theories?

Dieks et al. [2015, 209] claims that given the first commitment, "it is no longer clear that there exists an 'external' point of view that independently fixes the meanings of terms in the two theories". However, I must admit I do not see why this is the case: as I explained above, the first commitment only makes sense *if* we understand both theories as having pre-determined meanings, and *then* relating them via the duality/translation manual. In other words, the first commitment is perfectly compatible with the external view.

For the remainder of this paper I focus on the second commitment instead. I think the second commitment *does* entail that differentiating the two theories is meaningless, *only if* we believe that one should be a structural realist (epistemically/metaphysically) only about the empirical core of the duality, discarding the 'excess structure' which made the two theories distinct structures to begin with. In other words, we want to say that this 'excess structure' was not physically significant to begin with: only the empirical core was relevant to physics. It seems that this is required to make sense of the claim that it is 'meaningless' to say that one formulation, e.g. the **CFT** side, is better than the

other, e.g. the **AdS** side. If structural realism commits us only to the empirical core of the dual theories, then accordingly there is really only one structure in question. Hence, it is meaningless to ask which structure is better (there is only one). If there is only one structure, then the internal view seems to hold: under a structural realist view, the terms of the dual theories are defined in terms of their places in the structure. Hence, within the empirical core's structure, the different terms of the dual theories really mean the same thing, and hence we get some version of the internal view.

Why should we, even as structural realists, commit ourselves only to the empirical core? The argument seems to me to be an epistemic one: we should believe that the structure relevant to the two theories given the duality must really be common to both theories because, as De Haro claims above, "both theories are equally successful" by all epistemic criteria we apply. If this is true then it seems we have no way of differentiating between the two theories, and the best explanation for this epistemic equivalence is to appeal to their being 'the same' in some way. The only thing in common between the dual theories is the empirical core, so we should take this to be what explains their epistemic equivalence. Everything else (i.e. the 'excess structure') can be discarded, since they are irrelevant differences. As such, structural realism should commit us only to the empirical core.

However, it is not clear that the dual theories are indeed epistemically equivalent. In a naive sense, they are epistemically equivalent if one takes 'epistemic' to be 'empirical' equivalence. Given the duality, i.e. formal and predictive equivalence, it is trivial that the two theories are also 'empirically' equivalent. However, I do not think such a notion of empirical equivalence *exhausts* the epistemic criteria for differentiating between scientific theories. Of course, one main desideratum for scientific theorizing is to provide

Seattle, WA; 1-4 November 2018 -189-

predictions, descriptions and explanations of phenomena. Beyond that, though, I contend that another desideratum of scientific theorizing is to look for ways to develop better scientific theories, be it a more unificatory theory, a more explanatory theory, and so on.

We see this in play when De Haro discusses the position/momentum duality in quantum mechanics: "this duality is usually seen as teaching us something new about the nature of reality: namely, that atoms are neither particles, nor waves. By analogy, it is to be expected that gauge/gravity dualities teach us something about the nature of spacetime and gravity" [2017, 117]. However, this is only possible *if* the two theories were not epistemically equivalent! If they were epistemically equivalent, then how could we learn anything new from one theory that we cannot already learn from another? If 'area' and 'entanglement entropy' really meant the same thing and had the same physical interpretation, how could we learn something new when we realize that area can be related (via **RT**) to quantum entanglement? Indeed, this criticism extends generally to the internal view: how can we learn anything new from a duality if the dual theories are just the 'same theory', and indeed are *uninterpreted* to begin with? We learn something new when two *different* things are related in a surprising way, *especially* when they are related to other quantities, on each side, in interesting ways; I do not see how we can learn something new when one and the same thing is related to itself.

Furthermore, the two theories are *not* epistemically equivalent when we consider the methodological concerns of physicists, who generally note that the **CFT** is well-understood, while the dual string theory of gravity is not. For example, Horowitz and Polchinski [2009] notes that we only approximately understand the gravitational theory, but the **CFT** has been developed to very precise degrees. Lin points out that:

> A dictionary is reasonably well developed in the direction of using classical gravity to study the **CFT**, but the converse problem how to organize the information in certain **CFT**'s into a theory of quantum gravity with a semi-classical limit is hardly understood at all. [2015, 11]

If both theories are equally successful by *all* epistemic criteria we have, then this situation should not appear. Rather, it seems that scientific practice is of the opinion that the two theories are, in fact, *not* epistemically equal: one is more successful than the other in terms of a variety of criteria, such as precision of calculation, ease of understanding, availability of a non-perturbative analysis, and so on. It is one reason why **AdS/CFT** is such an interesting area of research: it allows us to understand a hard-to-understand theory in terms of an easier-to-understand theory. Unless one is given arguments for why such criteria should *not* be epistemically relevant, the dual theories, I contend, are *not* epistemically equivalent.

Of course, one could assume that the *goal* or *ideal*, when we fully understand the translation manual, is to render both theories equally epistemically successful. However, this presumes that both sides *will* end up being just as easy to compute, or understand, and so on. Of course, if we do discover a more fundamental characterization of *why* the two dual theories are related by the duality as such, e.g. the sort of 'deeper' theory Rickles [2011, 2017] hopes for, then clearly we are entitled to the internal view since this 'deeper' theory will ideally explain why the dual theories, despite their apparent differences, can be seen as different facets of a single theory, just like how special relativity unified electromagnetism and made it plausible to understand both the electric and magnetic fields as facets of the 'deeper' Faraday tensor field. Right now, though,

there is no such theory in sight, making this point inadequate for supporting the internal view.

Given the foregoing, it is not clear there is epistemic equivalence: the epistemic argument does not hold. The upshot is that we are not compelled to provide an explanation for why the dual theories are epistemically equivalent to begin with (they are not), and hence we have no need to commit ourselves only to the common empirical core, *even* as structural realists, nor to think that differentiating the dual theories is meaningless.

Recall the oddity I pointed out in §3, though. The claim of physical equivalence hangs on leaving the dual theories uninterpreted, but this latter claim was itself motivated by physical equivalence. It was hoped, then, that the epistemic argument could provide independent motivation for adopting physical equivalence. Given my criticism of De Haro's additional argument, though, the circle returns, and leaves the two claims uncompelling. Hence, we should not adopt the internal view itself. Furthermore, my criticisms suggest that the dual theories are in fact *not* epistemically equivalent, and this suggests that the default stance is one where the two theories are not theoretically equivalent at all. Given the duality, the only way this can be so is to adopt the view that the dual theories are physically non-equivalent; in other words we should adopt the external view instead.

To conclude, given the dialectic set up by the interpretive fork, and the inadequacies of the internal view, I suggest that we adopt the external view instead.

5 The Way Forward

Let me end by commenting on the external view and the broader debate on whether there is emergence given a duality (§1). In §3 we have seen how the internal view precludes emergence simply because there are no two distinct theories to speak of: we merely have two ways of looking at a single theory. This in turn swiftly rules out any talk of emergence. The external view, though, does not rule out emergence quite so easily, and there is some leeway to speak of emergence since we *do* have two distinct theories which are, as Teh noted, generically *not* isomorphic to one another. However, given the formal and predictive equivalences demanded by a duality relation, a duality relation is symmetric, and so there is nothing within a duality that will formally broker the asymmetry between two theories we often associate with emergence. One way to do so, as Teh (2013) suggests, is to introduce a claim of relative fundamentality, i.e. which theory is 'more fundamental' than another, is required to break the symmetry and provide us with the required asymmetry for emergence. While the external view does not entail this, it does not rule it out either. Hence, the external view does not preclude emergence; instead, it directs attention about emergence and duality away from the interpretative fork, onto whether and how one can make claims about relative fundamentality in the context of dualities. Alas, this requires much more attention than I can afford here: I leave it for another day.

References

Dieks, D., J. van Dongen, and S. D. Haro (2015). Emergence in Holographic Scenarios for Gravity. *Studies in the History and Philosophy of Modern Physics 52*, 203–216. 10.1016/j.shpsb.2015.07.007.

Fraser, D. (2017). Formal and Physical Equivalence in Two Cases in Contemporary Quantum Physics. *Studies in the History and Philosophy of Modern Physics 59*, 30–43. 10.1016/j.shpsb.2015.07.005.

Haro, S. D. (2017). Dualities and Emergent Gravity: Gauge/Gravity Duality. *Studies in the History and Philosophy of Modern Physics 59*, 109–125. DOI: 10.1016/j.shpsb.2015.08.004.

Haro, S. D., N. Teh, and J. Butterfield (2017). Comparing Dualities and Gauge Symmetries. *Studies in the History and Philosophy of Modern Physics 59*, 68–80. DOI: 10.1016/j.shpsb.2016.03.001.

Horowitz, G. and J. Polchinski (2009). Gauge/gravity duality. In D. Oriti (Ed.), *Approaches to quantum gravity: Toward a new understanding of space time and matter*, pp. 169–186. Cambridge: Cambridge University Press. arXiv:gr-qc/0602037.

Raamsdonk, M. V. (2010). Building up spacetime with quantum entanglement. *General Relativity and Gravitation 42*(10), 2323–2329. 10.1007/s10714-010-1034-0.

Rickles, D. (2011). A Philosopher Looks at Dualities. *Studies in the History and Philosophy of Modern Physics 42*(1), 54–67. DOI: 10.1016/j.shpsb.2010.12.005.

Rickles, D. (2017). Dual Theories: 'Same but Different or 'Different but Same? *Studies in the History and Philosophy of Modern Physics 59*, 62–67. 10.1016/j.shpsb.2015.09.005.

Ryu, S. and T. Takayanagi (2006). Holographic Derivation of Entanglement Entropy from AdS/CFT. *Phys. Rev. Lett 96*(18). 10.1103/PhysRevLett.96.181602.

Seiberg, N. (2007). Emergent spacetime. In D. Gross, M. Henneaux, and A. Sevrin (Eds.), *The Quantum Structure of Space and Time*, pp. 163–178. Singapore: World Scientific. DOI: 10.1142/9789812706768_0005.

Teh, N. (2013). Holography and Emergence. *Studies in the History and Philosophy of Modern Physics 44*(3), 300–311. DOI: 10.1016/j.shpsb.2013.02.006.

Causal Discovery and the Problem of Psychological Interventions

PSA 2018, Seattle

Markus Eronen

University of Groningen

m.i.eronen@rug.nl

Abstract

Finding causes is a central goal in psychological research. In this paper, I argue that the search for psychological causes faces great obstacles, drawing from the interventionist theory of causation. First, psychological interventions are likely to be both fat-handed and soft, and there are currently no conceptual tools for making causal inferences based on such interventions. Second, holding possible confounders fixed seems to be realistically possible only at the group level, but group-level findings do not allow inferences to individual-level causal relationships. I also consider the implications of these problems, as well as possible ways forward for psychological research.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -196-

1. Introduction

A key objective in psychological research is to distinguish causal relationships from mere correlations (Kendler and Campbell 2009; Pearl 2009; Shadish and Sullivan 2012). For example, psychologists want to know whether having negative thoughts is a cause of anxiety instead of just being correlated with it: If the relationship is causal, then the two are not just spuriously hanging together, and intervening on negative thinking is actually one way of reducing anxiety in patients suffering from anxiety disorders. However, to what extent is it actually possible to find psychological causes? In this paper, I will seek an answer this question from the perspective of state-of-the-art philosophy of science.

In philosophy of science, the standard approach to causal discovery is currently interventionism, which is a very general and powerful framework that provides an account of the features of causal relationships, what distinguishes them from mere correlations, and what kind of knowledge is needed to infer them (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009, Woodward 2003, 2015b; Woodward & Hitchcock 2003). Interventionism has its roots in Directed Acyclic Graphs (DAGs), also known as causal Bayes nets, which are graphical representations of causal relationships based on conditional independence relations (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009). More recently, James Woodward has developed interventionism into a full-blown philosophical account of causation, which has become popular in philosophy and the sciences. Several authors have also argued that interventionism adequately captures the role of causal thinking and reasoning in psychological research (Campbell 2007; Kendler and Campbell 2009; Rescorla forthcoming; Woodward 2008).

Seattle, WA; 1-4 November 2018 -197-

Based on interventionism, I will argue in this paper that the discovery of psychological causes faces great obstacles. This is due to problems in performing psychological interventions and deriving interventionist causal knowledge from psychological data.[1] Importantly, my focus is not on the existence or possibility of psychological causation, but on the *discovery* of psychological causes, which is a topic that has so far received little attention in philosophy.[2] Although I rely on interventionism, my arguments are based on rather general principles of causal inference and reasoning in science, and will thus apply to any other theory of causation that does justice to such principles.

The focus in this paper will be on the discovery *individual-level* (or within-subject) causes, not *population-level* (or between-subjects) causes. The first refers to causal relationships that hold for a particular individual: for example, John's negative thoughts cause John's problems of concentration. The latter refers to causal relationships that obtain in the population as a whole: for example, negative thoughts cause problems of concentration in a population of university students. It widely thought that ultimate goal of causal inference is to find individual-level causes, and that a population-level causal relationship should be seen as just an average of individual-level causal relationships (Holland 1986): For example, the causal relationship between negative thoughts and problems of concentration in a population of university students is only interesting insofar as it *also* applies to at least some of the individual students in the

[1] See Eberhardt (2013; 2014) for different (and domain-independent) problems for interventionist causal discovery.
[2] There is an extensive debate on the question whether interventionism vindicates non-reductive psychological causation by providing a solution the causal exclusion problem (e.g., Baumgartner 2009, Eronen 2012, Raatikainen 2010, Woodward 2015). I will sidestep this debate here, as my focus is not on the existence of non-reductive psychological causation, but on the discovery of psychological causes, be they reducible or not.

population.[3] Thus, in this paper I will discuss population-level causal relationships only when they are relevant to discovering individual-level causes.

Importantly, the distinction between population-level and individual-level causation is different from the distinction between type and token causation, even though the two distinctions are sometimes mixed up in the philosophical literature (see also Illari & Russo 2014, ch. 5). Token causation refers to causation between two actual events, whereas type causation refers to causal relationships that hold more generally. Individual-level causes can be either type causes or token causes. An example of an individual and type causal relationship would be that John's pessimistic thoughts cause John's problems of concentration: This is a general relationship between two variables, and not a relationship between two actual events. An example of an individual and token causal relationship would be that John's pessimistic thoughts before the exam on Friday at 2 pm caused his problems of concentration in the exam. As interventionism is a type-level theory of causation, and the aim of psychological research is primarily to discover regularities, not explanations to particular events, in this paper I will only discuss the discovery of type (individual) causes.

The structure of this paper is as follows. I will start by giving a brief introduction to interventionism, and then turn to problems of interventionist causal inference in psychology: First, to problems related to psychological interventions (section 2), and then to problems arising from the requirement to "hold fixed" possible confounders (section 3). After this, I will consider the possibility of the inferring psychological causes without interventions (section 4). In the last

[3] It has been argued that population-level (between-persons) causal relationships can also be real without applying to any individual (Borsboom, Mellenbergh, and van Heerden 2003). However, also those who believe in these kind of population-level causes agree that discovering individual causes is an important goal as well.

section, I discuss ways forward and various implications that my arguments have for psychology and its philosophy.

2. Interventionism

Interventionism is a theory of causation that aims at elucidating the role of causal thinking in science, and defining a notion of causation that captures the difference between causal relationships and mere correlations (Woodward 2003). Thus, the goal of interventionism is to provide a methodologically fruitful account of causation, and *not* to reduce causation to non-causal notions or analyse the metaphysical nature of causation (Woodward 2015b). In a nutshell, interventionist causation is defined as follows:

X is a cause of Y (in variable set **V**) if and only if it is possible to *intervene* on X to change Y when all other variables (in **V**) that are not on the path from X to Y are *held fixed* to some value (Woodward 2003).

Thus, in order to establish that X is a cause of Y, we need evidence that there is some way of intervening on X that results in a change in Y, when off-path variables are held fixed.[4] Importantly, it is not necessary to actually perform an intervention: What is necessary is knowledge on what *would* happen if we *were* to make the right kind of intervention.

[4] More precisely, this is the definition for a *contributing* cause. X is a *direct* cause of Y if and only if it is possible to intervene on X to change Y when all other variables (in **V**) are held fixed to some value (Woodward 2003). Thus, the definition of a contributing cause allows there to be other variables on the causal path between X and Y, whereas the definition of a direct cause does not. This does not reflect any substantive metaphysical distinction, as the question whether X is a direct or contributing cause is relative to what variables are included in the variable set. Importantly, notion of a contributing cause is *not* relative to a variable set in any strong sense – if X is a cause of Y in some variable set, then X will be a cause of Y in all variable sets where X and Y appear (Woodward 2008b). This is because the definition of an intervention is not relativized to a variable set.

The notion of an intervention plays a fundamental role in the account, and is very specifically defined. Here is a concise description of the four conditions that an intervention has to satisfy (Woodward 2003).

Variable *I* is an intervention variable for X with respect to Y if and only if:

(I1) *I causes* the change in X;

(I2) The change in X is *entirely* due to *I* and not any other factors;

(I3) *I* is not a cause of Y, or any cause of Y that is not on the path from X to Y;

(I4) *I* is *uncorrelated* with any causes of Y that are not on the path from X to Y.

The rationale behind these conditions is that if the intervention does not satisfy them, then one is not warranted to conclude that the change in Y was (only) due to the intervention on X. Thus, in simpler terms, the intervention should be such that it changes the value of the target variable *X* in such a way that the change in *Y* is *only* due to the change in *X* and not any other influences (Woodward 2015b). For example, if the intervention is correlated with some other cause of Y, say Z, that is not on the path from X to Y (violating I4), then the change in Y may have been (partly) due to Z, and not just due to X. Following standard terminology in the literature, I will call interventions that satisfy the criteria I1-I4 *ideal* interventions. I will now go through various problems in performing ideal interventions in psychology, starting from problems related to conditions I2 and I3 (section 3), and then turn to problems related to I4 and the "holding fixed" part of the definition of causation (section 4).

3. Psychological interventions

Seattle, WA; 1-4 November 2018 -201-

Before discussing psychological interventions, an important distinction needs to be made: The distinction between relationships where (1) the cause is *non-psychological*, and the *effect* is psychological, and (2) where the *cause* (and possibly also the effect) is *psychological*.[5] A large proportion (perhaps the majority) of experiments in psychology involve relationships of the first kind: The intervention targets a non-psychological variable (X) such as medication vs. placebo, therapy regime vs. no therapy, or distressing vs. neutral video, and the psychological effect of the manipulation of this non-psychological variable is tracked. In other words, the putative causal relation is between a non-psychological cause variable (X) and a psychological effect variable (Y). In these cases, it is possible to do (nearly) ideal interventions on the putative cause variable (X) by ensuring that the change in X was caused (only) by the intervention, that the intervention did not change Y directly, and that it was uncorrelated with other causes of Y. It is of course far from trivial to make sure that these conditions were satisfied, but as the variables intervened upon are non-psychological, making the right kinds of interventions is in principle not more difficult than in other fields. As regards the psychological effect variable (Y), there is no need to intervene on it; it is enough to measure the change in Y (which, again, is far from trivial, but faces just the usual problems in psychological measurement, which will be discussed below). The fact that many psychological experiments involve this kind of causal relationships may have contributed to the recent optimism on the prospects of interventionist causal inference in psychology.

[5] The line between psychological and non-psychological variables is likely to be blurry. However, for the present purposes it is not crucial where exactly the line should be drawn: My arguments apply to cases where it is clear that the cause variable is psychological (such as the examples in the main text), and such cases abound in psychological research.

However, psychological research also often concerns relationships of the second kind, that is, relationships where the *cause* is psychological. This is, for example, the case when the aim is to uncover psychological mechanisms that explain cognition and behavior (e.g., Bechtel 2008, Piccinini & Craver 2011), or to find networks of causally interacting emotions or symptoms (e.g., Borsboom & Cramer 2013). The reason why these relationships are crucially different from relationships of the first kind is that now the variable intervened upon is psychological, so the conditions on interventions now have to be applied to psychological variables.

Ideal interventions on psychological variables are rarely if ever possible. One reason for this has been extensively discussed by John Campbell (2007): Psychological interventions seem to be "soft", meaning that the value of the target variable *X* is not completely determined by the intervention (Eberhardt & Scheines 2007; see also Kendler and Campbell 2009; Korb and Nyberg 2006). In other words, the intervention does not "cut off" all causal arrows ending at *X*. As a non-psychological example, when studying shopping behaviour during one month by intervening on income, an ideal intervention would fully determine the exact income that subjects have that month, whereas simply giving the subjects an *extra* 5000€ would count as a soft intervention (Eberhardt & Scheines 2007). Similarly, if we intervene on John's psychological variable *alertness* by shouting "WATCH OUT!", this does not completely cut off the causal contribution of other psychological variables that may influence John's *alertness*, but merely adds something on top of those causal contributions (Campbell 2007). As most or all interventions on psychological variables are likely to be soft, Campbell proposes that we should simply allow such soft interventions in the context of psychology. Campbell argues that these kind of interventions can still be informative and indicative of causal relationships (Campbell

2007), and this conclusion is supported by independent work on soft interventions in the causal modelling literature (e.g., Eberhardt & Scheines 2007; Korb and Nyberg 2006).

However, the problem of psychological interventions is not solved by allowing for soft interventions. There is a further, equally important reason why interventions on psychological variables are problematic: Psychological interventions typically *change several variables simultaneously*. For example, suppose we wanted to find out whether *pessimistic thoughts* cause *problems in concentration*. In order to do this, we would have to find out what would happen to *problems in concentration* if we were to intervene just on *pessimistic thoughts* without perturbing other psychological states with the intervention. However, how could we intervene on *pessimistic thoughts* without changing, for example, *depressive mood* or *feelings of guilt*? As an actual scientific example, consider a network of psychological variables that includes, among others, the items *alert*, *happy*, and *excited* (Pe et al. 2015). How could we intervene on just one of those variables without changing the others?

One reason why performing "surgical" interventions that only change one psychological variable is so difficult is that there is no straightforward way of manipulating or changing the values of psychological variables (as in, for example, electrical circuits). Interventions in psychology have to be done, for example, through verbal information (as in the example of John above) or through visual/auditory stimuli, and such manipulations are not precise enough to manipulate just one psychological variable. Also state-of-the-art neuroscientific methods such as Transcranial Magnetic Stimulation affect relatively large areas of the brain, and are not suited for intervening on specific psychological variables. Currently, and in the foreseeable future, there is no realistic

way of intervening on a psychological variable without at the same time perturbing some other psychological variables.

Thus, it is likely that most or even all psychological interventions do not just change the target variable X, but also some other variable(s) in the system. In the causal modelling literature, interventions of this kind have been dubbed *fat-handed*[6] interventions (Baumgartner and Gebharter 2016; Eberhardt & Scheines 2007; Scheines 2005). For example, an intervention on pessimistic thoughts that also immediately changes depressive mood is fat-handed. Fat-handed interventions have been recently discussed in philosophy of science, but mainly in the context of mental causation and supervenience (e.g., Baumgartner and Gebharter 2016, Romero 2015), and the fact that psychological interventions are likely to be systematically fat-handed (for reasons unrelated to supervenience) has not yet received attention.

An additional complication is that it is difficult check what a psychological intervention precisely changed, and to what extent it was fat-handed (and soft). In fields such as biology or physics there are usually several independent ways of measuring a variable: for example, temperature can be measured with mercury thermometers or radiation thermometers, and the firing rate of a neuron can be measured with microelectrodes or patch clamps. However, measurements of psychological variables, such as emotions or thoughts, are based on self-reports, and there is no further independent way of verifying that these reports are correct. Moreover, only a limited number of psychological variables can be measured at a given time point, so an intervention may always have unforeseen effects on unmeasured variables.

[6] According to Scheines (2005), this term was coined by Kevin Kelly.

Seattle, WA; 1-4 November 2018 -205-

Why are fat-handed interventions so problematic for interventionist causal inference? The reason becomes clear when looking at condition I3: The intervention should not change any variable Z that is on a causal pathway that leads to Y (except, of course, those variables that are on the path between X and Y). If the causal structure of the system under study is known, as well as the changes that the intervention causes, then this condition can sometimes be satisfied even the intervention was fat-handed. However, in the context of intervening on psychological variables, neither the causal structure nor the exact effects of the interventions are known. Thus, when the intervention is fat-handed, it is not known whether I3 is satisfied or not, and in many cases it is likely to be violated. In other words, we cannot assume that the intervention was an unconfounded manipulation of X with respect to Y, and cannot conclude that X is a cause of Y.

4. The Problem of "Holding Fixed"

The next problem that I will discuss is related to the last part of the definition of interventionist causation: X is a cause of Y (in variable set **V**) if and only if it is possible to intervene on X to change Y *when all other variables (in **V**) that are not on the path from X to Y are held fixed to some valu*e. The motivation for this requirement is to make sure that the change in Y is really due to the change X, and not due to some other cause of Y. To a large extent, this is just another way of stating what is already expressed in the definition of an intervention, in conditions I3 and I4: The intervention should not be confounded by any cause of Y that is not on the path between X and Y.[7] In the previous section, we saw that fat-handed interventions pose a challenge for

[7] In recent publications, Woodward often gives a shorter definition of causation that does not include the "holding fixed" part, for example: "X causes Y if and only if under some interventions on X (and possibly other variables) the value of Y changes" (Woodward 2015). This is understandable, as the definition of intervention already contains conditions I3 and I4, which effectively imply holding fixed potential causes of Y that are correlated with the intervention and are not on the path from X to Y. However, there are also good reasons why the full definition has to

satisfying this condition. However, as I will now show, it is problematic in psychology also for more general reasons.

In psychology, it is impossible to hold psychological variables fixed in any concrete way: We cannot "freeze" mental states, or ask an individual to hold her thoughts constant. Thus, the same effect has to be achieved indirectly, and the gold standard for this is Randomized Controlled Trials (RCTs) (Woodward 2003, 2008). RCTs have their origin in medicine, but are widely used in psychology and the social sciences as well (Clarke et al. 2014; Shadish, Cook and Campbell 2002; Shadish and Sullivan 2012). The basic idea of RCTs is to conduct a trial with two groups, the test group and the control group, which are as similar to one another as possible, but the test group receives the experimental manipulation and the control group does not. If the groups are large enough and the randomization is done correctly, any differences between the groups should be only due to the experimental manipulation. If everything goes well, this in effect amounts to "holding fixed" all off-path variables.

However, this methodology has an important limitation that has been overlooked in the literature on interventionism. As the effect of "holding fixed" is based on the difference between the groups as wholes, it only applies at the level of the group, and not at the level of individuals. For this reason, results of RCTs hold for the study population as a whole, but not necessarily for particular individuals in the population (cf. Borsboom 2005, Molenaar & Campbell 2009). For example, if we discover that pessimistic thoughts are causally related to problems of

include the second component as well. For example, consider a situation where we intervene on X with respect to Y, and Y changes, but this change is fully due to a change in variable Z, which is a cause of Y that is *uncorrelated* with the intervention variable. In this situation, without the "holding fixed" requirement we would falsely conclude that X is a cause of Y.

Seattle, WA; 1-4 November 2018 -207-

concentration in the population under study, it does not follow that this causal relationship holds in John, Mary, or any other specific individual in the population. This is related to the "fundamental problem of causal inference" (Holland 1986): Each individual in the experiment can belong to only one of the two groups (control or test group), and therefore cannot act as a "control" for herself, so only an average causal effect can be estimated. What this implies for causal inference in psychology is that when a causal relationship is discovered through an RCT, we cannot infer that this relationships holds for any specific individual in the population (see also Illari & Russo 2014, ch. 5).

This does not mean that the population-level findings based on RCTs are uninformative or useless. The point is rather that we currently have no understanding of when, to what extent and under what circumstances they also apply to the individuals in the population. This of course applies also to other fields where RCTs are used, such the biomedical sciences. Indeed, especially in the context of personalized medicine, the fact that RCTs are as such not enough to establish individual-level causal relationships has recently become a matter of discussion (e.g., de Leon 2012).

It might be tempting to simply look at the data more closely and find those individuals for whom the intervention on X actually corresponded with a change in Y. However, it would be a mistake to conclude that in those individuals the change in Y was caused by X. It might very well have been caused by some other cause of Y, as possible confounders were only held fixed at the group level, not at the individual level.[8] Thus, in RCTs possible confounders can only be held fixed at

[8] Would it be possible for a causal relationship to hold at the population level, but not for any individual in the population? Probably not, if the relationship is genuine: Weinberger (2015) has argued that there has to be at least *one* individual in the population for whom the relationships holds. However, in the context of discovery, it is

the group level, and this does not warrant causal inferences that apply to specific individuals. This is further limitation to interventionist causal inference in psychology.

5. Finding psychological causes without interventions

One possible response to the concerns raised in the previous two sections is that interventionism does not require that interventions are actually performed: As briefly mentioned in section 2, what is necessary is to know what *would* happen if we *were* to perform the right kinds of interventions. In other words, in order to establish that X is a cause of Y, it is enough to know that if we *were* to intervene on X with respect to Y (while holding off-path variables fixed), then Y *would* change. For example, it is beyond doubt that the gravitation of the moon causes the tides, even though no one has ever intervened on the gravitation of the moon to see what happens to the tides, and such an intervention would be practically impossible (Woodward 2003). Similarly, it could be argued that even though it is practically impossible to do (ideal) interventions on psychological variables, the knowledge on the effects of interventions could be derived in some other way. Let us thus consider to what extent this could be possible.

The state-of-the-art method for deriving (interventionist) causal knowledge when data on interventions is not available is *Directed Acyclic Graphs (DAGs)*, which were briefly mentioned in the introduction (see also Malinsky & Danks 2018, Pearl 2000, Spirtes, Glymour and Scheines 2000, Spirtes & Zhang 2016). Causal discovery algorithms based on DAGs take purely

possible that a causal *finding* at the population level is just an artefact of heterogeneous causal structures at the individual level, and therefore does not apply to any individual in the population.

observational data as input, and based on conditional independence relations, find the causal graph that best fits the data. In principle, these algorithms can be used for psychological data, with the aim of discovering causal relationships between psychological variables.

However, even though these algorithms do not require experimental data, they do require data from which conditional independence relations can be reliably drawn, and they (implicitly) assume that the variables that are modelled are independently and surgically manipulable (Malinsky & Danks 2018). In contrast, as should be clear from the above discussion, measurements of psychological variables typically come with a great deal of uncertainly, and it is not clear to what extent they can be independently manipulated. Moreover, causal discovery algorithms standardly assume *causal sufficiency*, that is, that there are no unmeasured common causes that could affect the causal structure (Malinsky & Danks 2018; Spirtes & Zhang 2016). The reason for this is that if two or more variables in the variable set have unmeasured common causes, then the inferences concerning the causal relationships between those variables will be either incorrect or inconclusive. However, missing common causes is likely the norm rather than the exception when it comes to psychological variables. For example, if the variable set consists of, say, 16 emotion variables, how likely is it that *all* relevant emotion variables have been included? And even if all emotion variables that are common causes to other emotion variables are included, is it plausible to assume that there are no further cognitive or biological variables that could be common causes to some of the emotion variables? As similar questions can be asked for any context involving psychological variables, causal sufficiency is a very unrealistic assumption for psychological variable sets.

For these reasons, psychological data sets are rather ill-suited for causal discovery algorithms, and these algorithms cannot be treated as reliable guides to interventionist causal knowledge in psychology. It is likely that the problems of psychological interventions discussed in the previous sections are not just practical problems in carrying out interventions, but reflect the immense complexity of the system under study (the human mind-brain), and therefore cannot be circumvented by just using non-experimental data (see, however, section 7 for a different approach).

6. Psychological interventions: A summary

To summarize, what I have argued so far is that interventionist causal inference in psychology faces several obstacles: (1) Psychological interventions are typically *both* fat-handed *and* soft: They change several variables simultaneously, and do not completely determine the value(s) of the variable(s) intervened upon. It is not known to what extent such interventions give leverage for causal inference. (2) Due to the nature psychological measurement, the degree to which a psychological intervention was soft and fat-handed, or more generally, what the intervention in fact did, is difficult to reliably estimate. (3) Holding fixed possible confounders is only possible at the population level, not at the individual level, and it is not known under what conditions population-level causal relationships also apply to individuals. (4) Causal inference based on data without interventions requires assumptions that are unrealistic for psychological variable sets. Taken together, these issues amount to a formidable challenge for finding psychological causes.[9]

[9] Baumgartner (2009, 2012, 2018) has argued that mental-to-physical supervenience makes it impossible to satisfy the Woodwardian conditions on interventions, and that if interventionism is modified to accommodate supervenience relationships (as in Woodward 2015), the result is that any causal structure with a psychological

Seattle, WA; 1-4 November 2018 -211-

7. Discussion

Although various metaphysical and conceptual issues related to psychological causation have been extensively discussed in philosophy of science, little attention has been paid to the *discovery* of psychological causes. In this paper, I have contributed to filling this lacuna, by discussing the search for psychological causes in the framework of the interventionist theory of causation. The upshot is that finding individual psychological causes faces daunting challenges. The problems in holding fixed confounders and performing interventions need to be taken into account when trying to establish a psychological causal relationship, or when making claims about psychological causes.

However, I do not want to argue that finding psychological causes is *impossible*, or that researchers should stop looking for psychological causes. Rather, my aim is to contribute to getting a better understanding of the limits of finding causes in psychology, and the challenges involved. This can also lead to positive insights regarding causal inference in psychology. One such insight is that more attention should be paid to *robust inference* or *triangulation*. Often when individual methods or sources of evidence are insufficient or unreliable, as is the case here, what is needed is a more holistic approach. A widespread (though not uncontroversial) idea in philosophy of science is that evidence from several independent sources can lead to a degree of confidence even if the sources are individually fallible and insufficient (Eronen 2015, Kuorikoski

cause becomes empirically indistinguishable from a corresponding structure where the psychological variable is epiphenomenal. If this reasoning is correct, it leads to a further (albeit more theoretical) problem for interventionist causal inference: Any empirical evidence for a causal relationships with a psychological cause is equally strong evidence for a corresponding epiphenomenal structure, and it is not clear which structure should be preferred and on what grounds.

& Marchionni 2017, Munafo & Smith 2017, Wimsatt 1981, 1994/2007). For example, there is no single method or source of evidence that would be individually sufficient to establish that the anthropogenic increase in carbon dioxide is the cause for the rise in global temperature, but there is so much converging evidence from many independent sources that scientists are confident that this causal relationship exists. Similarly, evidence for a psychological causal relationship could be gathered from many independent sources: Several different (soft and fat-handed) interventions involving different variables, multilevel models based on time-series data, single-case observational studies, and so on.[10] If they all point towards the same causal relationships, this may lead to a degree of confidence in the reality of that relationship. However, how this integration of evidence would exactly work, and whether it can actually lead to sufficient evidence for psychological causal relationships, are open questions.

A related point is that psychological research can also make substantive progress *without* establishing causal relationships. Often important discoveries in psychology have not been discoveries of causal relationships, but rather discoveries of robust *patterns* or *phenomena* (Haig 2012, Rozin 2001, Tabb and Schaffner 2017). Consider, for example, the celebrated discovery that people often do not reason logically when making statistical predictions, but rely on shortcuts, for example, grossly overestimating the likelihood of dying in an earthquake or terror attack (Kahneman & Tversky 1973). In other words, when we reason statistically, we often rely on heuristics that lead to biases. The discovery of this phenomenon had nothing to with methods of causal inference (Kahneman and Tversky 1973), and its significance is not captured by describing causal relationships between variables. In fact, the causal mechanisms underlying the

[10] See also Peters, Bühlmann, & Meinshausen (2016), who present a formal model for inferring causal relationships based on their stability under different kinds of (non-ideal) interventions.

Seattle, WA; 1-4 November 2018 -213-

heuristics and biases of reasoning are still unknown. Similar examples abound in psychology: Consider, for example, groupthink or inattentional blindness. Of course, there are likely to be causal mechanisms that give rise to these phenomena, but the phenomena are highly relevant for theory and practice even when we know little or nothing about those underlying mechanisms (which is the current situation). This, in combination with the challenges discussed in this paper, suggests that (philosophy of) psychology might benefit from reconsidering the idea that discovering causal relationships is central for making progress in psychology.

Finally, one might wonder whether the problems I have discussed here are restricted to just psychology. Indeed, I believe that the arguments I have presented are more general, and apply to any other fields where there are similar problems with soft and fat-handed interventions and controlling for confounders. There is probably a continuum, where psychology is close to one end of the continuum, and at the other end we have fields where ideal interventions can be straightforwardly performed and variables can be easily held fixed, such as engineering science. Fields such as economics and political science are probably close to where psychology is, as they also face deep problems in making (ideal) interventions and measuring their effects. Same holds for neuroscience, at least cognitive neuroscience: The problems of soft and fat-handed interventions and holding variables fixed apply just as well to brain areas as to psychological variables (see also Northcott forthcoming). Thus, appreciating the challenges I have discussed here and considering possible reactions to them could also benefit many other fields besides psychology.

To conclude, I have argued in this paper that there are several serious obstacles to the discovery of psychological causes. As it is widely assumed in both psychology and its philosophy that the discovery of causes is a central goal, these obstacles need to be explicitly discussed, taken into account, and studied further.

References

Baumgartner, M. (2013). Rendering Interventionism and Non-Reductive Physicalism Compatible. *dialectica* 67: 1-27.

Baumgartner, M. (2018). The Inherent Empirical Underdetermination of Mental Causation. *Australasian Journal of Philosophy.*

Baumgartner, M and Gebharter, A. (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science* 67: 731-756.

Borsboom, Dennny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press

Borsboom, Denny and Anqelique O. Cramer. 2013. "Network analysis: an integrative approach to the structure of psychopathology." *Annual review of clinical psychology* 9: 91-121.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203.

Campbell, John. 2007. "An interventionist approach to causation in psychology." In: A. Gopnik & L. Schulz (eds.) *Causal Learning. Psychology, Philosophy, and Computation*. Oxford: Oxford University Press, 58–66.

Chirimuuta, Mazviita. Forthcoming. “Explanation in Computational Neuroscience: Causal and Non-causal.” *British Journal for the Philosophy of Science*. DOI:https://doi.org/10.1093/bjps/axw034

Clarke, B., Gillies, D., Illari, P, Russo, F., & Williamson, J. 2014. “Mechanisms and the evidence hierarchy.” *Topoi* 33: 339-360.

de Leon, J. (2012). Evidence-based medicine versus personalized medicine: are they enemies? *Journal of clinical psychopharmacology*, *32*(2), 153-164.

Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science*, *80*(5), 684-696.

Eberhardt, F. (2014). Direct causes and the trouble with soft interventions. *Erkenntnis*, *79*(4), 755-777.

Eberhardt, Frederick and Richard Scheines. 2007. “Interventions and causal inference.” *Philosophy of Science* 74: 981–995.

Eronen, Markus. Forthcoming. “Interventionism for the Intentional Stance: True Believers and Their Brains.” *Topoi*.

Hamaker, Ellen L. 2011. “Why researchers should think **“**within-person.**””** In M. R. Mehl, & T. A. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Kahneman, Daniel and Amos Tversky. 1973. “On the psychology of prediction.” *Psychological Review* 80: 237-251.

Kendler, Kenneth S. and John Campbell. 2009. Interventionist causal models in psychiatry: repositioning the mind-body problem. *Psychological Medicine* 39: 881-887.

Korb, K. B., & Nyberg, E. 2006. "The power of intervention." *Minds and Machines* 16*:* 289-302.

Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science, 83*, 227-247.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470.

Menzies, Peter. 2008. "The exclusion problem, the determination relation, and contrastive causation." In J. Hohwy & J. Kallestrup (Eds.) *Being Reduced* (pp. 196-217). Oxford: Oxford University Press.

Molenaar, Peter and Cynthia Campbell. 2009. "The new person-specific paradigm in psychology." *Current Directions in Psychological Science* 18: 112-117.

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature* 553, 399-401

Northcott, R. (forthcoming). Free will is not a testable hypothesis. *Erkenntnis*.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... & Kuppens, P. 2015. "Emotion-network density in major depressive disorder." *Clinical Psychological Science*, *3*(2), 292-300.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, Judea. 2009. "Causal inference in statistics: An overview." *Statistics surveys* 3: 96-146.

Pearl, Judea. 2014. "Comment: understanding simpson's paradox." *The American Statistician* 68: 8-13.

Peters, J. , Bühlmann, P. and Meinshausen, N. (2016), Causal inference by using invariant prediction: identification and confidence intervals. J. R. Stat. Soc. B, 78: 947-1012. doi:10.1111/rssb.12167

Rescorla, Michael. Forthcoming. “An interventionist approach to psychological explanation.” *Synthese*.

Reutlinger, Alexander and Juha Saatsi (eds.). 2017. *Explanation Beyond Causation*. Oxford: Oxford University Press.

Romero, F. (2015). Why there isn’t inter-level causation in mechanisms. *Synthese*, *192*(11), 3731-3755.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2-14.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science*, *72*(5), 927-940.

Shadish W. R., Cook T. D. and Campbell D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin; Boston.

Shadish, W. R., & Sullivan, K. J. 2012. “Theories of causation in psychological science.” In H. Cooper et al. (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 23-52). Washington, DC: American Psychological Association.

Shapiro, Lawrence. 2010. “Lessons from causal exclusion.” *Philosophy and Phenomenological Research*, 81, 594-604.

Shapiro, Lawrence. 2012. “Mental manipulations and the problem of causal exclusion.” *Australasian Journal of Philosophy*, 90, 507–524.

Shapiro, Lawrence and Elliott Sober. 2007. “Epiphenomenalism: the dos and the don’ts.” In G. Wolters & P. Machamer (Eds.) *Thinking about causes: from Greek philosophy to modern physics* (pp. 235–264). Pittsburgh, PA: University of Pittsburgh Press.

Spirtes, Peter, Glymour, Clark and Richard Scheines. 2000. *Causation, prediction, and search*. New York: Springer.

Tabb, K., & Schaffner, K. F. (2017). Causal pathways, random walks and tortuous paths: Moving from the descriptive to the etiological in psychiatry. In: Kendler, K. S., & Parnas, J. (Eds.) *Philosophical Issues in Psychiatry IV: Nosology* (pp. 342-360). Oxford: Oxford University Press.

Weinberger, Naftali. 2015. “If intelligence is a cause, it is a within-subjects cause.” *Theory & Psychology*, 25(3), 346-361.

Woodward, James. 2003. *Making things happen. A theory of causal explanation.* Oxford: Oxford University Press.

Woodward, James. 2008. “Mental causation and neural mechanisms.” In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: new essays on reduction, explanation, and causation*. Oxford: Oxford University Press: 218-262

Woodward, James. 2015a. “Interventionism and causal exclusion.” *Philosophy and Phenomenological Research* 91, 303-347.

Woodward, James. 2015b. “Methodology, ontology, and interventionism.” *Synthese* 192, 3577-3599.

Woodward, James & Christopher Hitchcock. 2003. “Explanatory Generalizations, Part I: A Counterfactual Account.” *Noûs* 37(1): 1–24.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

Why Replication is Overrated

Current debates about the replication crisis in psychology take it for granted that direct replication is valuable and focus their attention on questionable research practices in regard to statistical analyses. This paper takes a broader look at the notion of replication as such. It is argued that all experimentation/replication involves individuation judgments and that research in experimental psychology frequently turns on probing the adequacy of such judgments. In this vein, I highlight the ubiquity of conceptual and material questions in research, and I argue that replication is not as central to psychological research as it is sometimes taken to be.

1. Introduction: **The "Replication Crisis"**

In the current debate about replicability in psychology, we can distinguish between (1) the question of why not more replication studies are done (e.g., Romero 2017) and (2) the question of why a significant portion (more than 60%) of studies, when they *are* done, fail to replicate (I take this number from the Open Science Collaboration, 2015). Debates about these questions have been dominated by two assumptions, namely, first, that it is in general desirable that scientists conduct replication studies that come as close as possible to the original, and second, that the low replication rate can often be attributed to statistical problems with many initial studies, sometimes referred to as "p-hacking" and "data-massaging."[1]

[1] An important player in this regard is the statistician Andrew Gelman who has been using his blog as a public platform to debate methodological problems with mainstream social psychology (http://andrewgelman.com/).

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

I do not wish to question that close (or "direct") replications can sometimes be epistemically fruitful. Nor do I wish to question the finding that are severe problems in the statistical analyses of many psychological experiments. However, I contend that the focus on formal problems in data analyses has come at the expense of questions about the notion of *replication* as such. In this paper I hope to remedy this situation, highlighting in particular the implications of the fact that psychological experiments in general are infused with conceptual and material presuppositions. I will argue that once we gain a better understanding of what this entails with respect to replication, we get a deeper appreciation of philosophical issues that arise in the investigative practices of psychology. Among other things, I will show that replication is not as central to these practices as it is often made out to be.

The paper has three parts. In part 1 I will briefly review some philosophical arguments as to why there can be no exact replications and, hence, why attempts to replicate always involve individuation judgments. Part 2 will address a distinction that is currently being debated in the literature, i.e., that between direct and conceptual replication, highlighting problems and limitations of both. Part 3, finally, will argue that a significant part of experimental research in psychology is geared toward exploring the shape of specific phenomena or effects, and that the type of experimentation we encounter there is not well described as either direct or conceptual replication.

2. The Replication Crisis and the Ineliminability of Concepts

When scientists and philosophers talk about successfully replicating an experiment, they typically mean that they performed the same experimental operations/interventions. But what does it mean to perform "the same" operations as the ones performed by a previous experiment? With regard to this question, I take it to be trivially true that two experiments cannot be identical: At the very least, the time variable will differ. Replication can therefore at best aim for *similarity* (Shavit & Ellison 2017), as is also recognized by some authors in psychology. In this vein, Lynch et al (2015) write that "[e]xact

replication is impossible" (Lynch et al 2015, 2), arguing that at most advocates of direct replication can aim for is to get "as close as possible," i.e., to conduct an experiment that is similar to the previous one. In the literature, such experiments are also referred to as "direct replications." (e.g., Pashler & Harris 2912).[2]

The notion of similarity is, of course, also notoriously problematic (e.g., Goodman 1955), since any assertion of similarity between A and B has to specify with regard to what they are similar. In the context of experimentation, the relevant kinds of specifications already presuppose conceptual and material assumptions, many of which are not explicated, about the kinds of factors one is going to treat as relevant to the subject matter (see also Collins 1985, chapter 2). Such conceptual decisions will inform what one takes to be the "experimental result" down the line (Feest 2016). For example, If I am interested in whether listening to Mozart has a positive effect on children's IQ, I will design an experiment, which involves a piece by Mozart as the independent variable and the result of a standardized IQ-test at a later point. Now if I get an effect, and if I call it a Mozart effect, I am thereby assuming that the piece of music I used was causally responsible *qua being a piece by Mozart*. Moreover, when I claim that it's an effect on intelligence, I am assuming that the test I used at the end of the experiment *in fact measured intelligence*. These judgments rely on conceptual assumptions already built into the experiment qua choice of independent and dependent variables. In addition, I need *material assumptions* to the effect that potentially confounding variables have been controlled for. I take this example to show that whenever we investigate an effect *under a description*, we cannot avoid making conceptual assumptions when determining whether an experiment has succeeded or failed. This goes for original experiments as well as for replications.

[2] Both advocates and critics of direct replication sometimes contrast such replications with "conceptual" replications" (Lynch et al 2015). We will return to this distinction below.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

One obvious rejoinder to this claim might be to say that replication attempts need not investigate effects under a description. They might simply imitate what the original experiment did, with no particular commitment to what is being manipulated or measured. But even if direct replications need not explicitly replicate effects under a description, I argue that they nonetheless have to make what Lena Soler calls "individuation judgments" (Soler 2011). For example, the judgment that experiment 2 is relevantly similar to experiment 1 involves the judgment that experiment 2 does not introduce any confounding factors that were absent in experiment 1. However, such judgments have to rely on some assumptions about what is relevant and what is irrelevant to the experiment, where these assumptions are often unstated auxiliaries. For example, I may (correctly or incorrectly) tacitly assume that temperature in the lab is irrelevant and hence ignore this variable in my replication attempt.

It is important to recognize that the individuation judgments made in experiments have a high degree of epistemic uncertainty. Specifically, I want to highlight what I call the problem of "conceptual scope," which arises from the question of how the respective independent and dependent variables are described. Take, for example, the above case where I play a specific piece by Mozart in a major key at a fast pace. A lot hangs on what I take to be the relevant feature of this stimulus: the fact that it's a piece by Mozart, the fact that it's in a major key, the fact that it's fast? etc. Depending on how I describe the stimulus, I might have different intuitions about possible confounders to pay attention to. For example, if I take the fact that a piece is by Mozart as the relevant feature of the independent variable, I might control for familiarity with Mozart. If I take the relevant feature to be the key, I might control for mood. Crucially, even though scientists make decisions on the basis of (implicit or explicit) assumptions about conceptual scope, their epistemic situation is typically such that they don't know what is the "correct" scope. This highlights a feature of psychological experiments that is rarely discussed in the literature about the replication crisis, i.e., the deep epistemic uncertainty and conceptual openness of much

Seattle, WA; 1-4 November 2018 -223-

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

research. This concerns both the initial and the replication study. Thus, concepts are ineliminable in experimental research, while at the same time being highly indeterminate.

3. Is the dichotomy between direct and less direct replication pragmatically useful?

One way of paraphrasing what was said above is that all experiments involve individuation judgments and that this concerns both original and replication studies. While this serves as a warning against a naïve reliance on direct (qua non-conceptual) replication, it might be objected that direct replications nonetheless make unique epistemic contributions. This is indeed claimed by advocates of both direct and less direct (="conceptual") replication alike. I will now evaluate claims that have aligned the distinction between direct and "conceptual" with some relevant distinctions in scientific practice, such as that between the aim of establishing the existence of a phenomenon and that of generalizing from such an existence claim on the one and that between reliability and validity on the other. I will argue that while these distinctions are heuristically useful, but on closer inspection bring to the fore exactly the epistemological issues just discussed.

3.1 Existence vs. Generalizability

Many scientists take it as given that there cannot be two identical experiments, but nonetheless argue that there is significant epistemic merit in trying to get *close enough*., i.e., to conduct direct replications. In turn, the notion of a direct replication is frequently contrasted with that of a "conceptual" replication. In a nutshell, direct replications essentially try to redo "the same" experiment (or at least something very close), whereas the conceptual replications try to operationalize the same question or concept/effect in a different way. The advantage of direct replications, as viewed by its advocates, is that by being able to redo an experiment faithfully and to create the same effect, one can show that the effect was real: "Exact and very close replications establish the basic existence and stability of a

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -224-

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

phenomenon by falsifying the (null) hypothesis that observations simply reflect random noise" (LeBel et al, forthcoming, 7).

Advocates of conceptual replication don't deny this advantage of close replications, but hold that we want more than to establish that a given effect – created under very specific experimental conditions – is real. We want to know whether our findings about it can be generalized to: "When the goal is generalization, we argue that 'imperfect' conceptual replications that stretch the domain of research may be more useful" (Lynch et al 2015, 2). From a strictly Popperian perspective, the idea that non-falsification of the hypothesis of random error can provide proof of stability and existence is questionable , of course. But even if we abandon Popperian ideology here and take the falsification of H0 (that the initial effect was due to random error) to point to the truth of H1 (that there is a stable effect), the question is how to describe the effect. In other words, when claiming to have confirmed an effect, we have to say *what kind of effect* it is. And there we face the following dilemma:

a) Either we describe the effect as highly specific to very local experimental circumstances, involving the choice of a specific independent variable, delivered in a specific way etc.
b) Or we describe it in slightly broader terms, e.g., as a Mozart effect.

Advocates of direct replication might indeed endorse something like a), thereby exhibiting the kind of caution that motivated early operationists, in that no claim is made beyond the confines of a specific experiment. If, on the other hand, psychologists endorsed a description such as b), they would immediately run into the question of conceptual scope, i.e., the question *under what description* the independent variable can be said to have caused an effect. I argue that no amount of direct replication can answer this question, and hence, even if direct replication can confirm the existence of an effect, it cannot say what kind of effect. By asserting this, I am not saying that it's never useful to do a direct replication. My claim is merely that it will tell us relatively little. More pointedly: Direct replication can (perhaps) provide evidence for the existence of something, but it cannot say *existence of what*. Rolf

Zwaan makes a similar point when he states that "replication studies "tell us about the reliability of those findings. They don't tell us much about their validity." (Zwaan 2013).

In a similar vein, I argue that direct replication, with its narrow focus on ruling out random error, is epistemically unproductive, because it has nothing to say about *systematic error*. Systematic error arises if one erroneously attributes an effect to a specific feature of the experiment, when it is in fact due to another feature of the experiment. This can include, but is not limited to, the above-mentioned problem of conceptual scope. Fiedler et al. (2012) make a similar point when they argue that a narrow focus on falsification (with the aim of avoiding false positives) can be detrimental to the research process. Differently put, by privileging direct replication, we are not in a position to inquire about the kind of effect in question. This question, I argue, is best addressed by paying close attention to the possibility of systematic error, and hence by doing conceptual work. In other words, experimentally probing into systematic errors of conceptual scope is a valuable and productive part of the research process as it enables scientists to gradually explore what kind of effect (if any) they are looking at.[3]

3.2 Generality

I have argued that (a) scientists typically produce effects under a description and (b) that it can be epistemically productive to probe the scope of the description and to investigate the possibility of systematic error with regard to experiments that draw on such descriptions. It is epistemically productive, because it forces scientists to explore the nature and boundaries of the effect they are investigating. With this I have argued against a narrow focus on direct replication and I have cautioned against overstating the epistemic merits of such replication. But when we are concerned with effects

[3] I take this to be a contribution to arguments that philosophers of experimentation have made for a long time; e.g., Mayo 1996.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -226-

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

under a description, we are confronted with questions about the adequacy of the description. It is this question that advocates of "conceptual replication" claim to be able to address when they emphasize that their approach can deliver generality (over mere existence).

We have to distinguish between two notions of generality, namely (a) what kinds of descriptions one can generalize or infer to *within the experiment*, and (b) does the effect in question hold *outside the lab* (see Feest & Steinle 2016). These types of generality are also sometimes referred to as internal vs. external validity, respectively (Campbell & Stanley 1966; Guala 2012), where the former refers to the quality of inferences within an experiment and the latter refers to the quality of inferences from a lab to the world. The notion of generalizability raises questions about two kinds of validity. My focus here will be on internal validity, i.e., with the question of whether the effect generated in an experiment really exists as described by the scientist.[4]

Internal validity can fail to hold because of epistemic uncertainties regarding confounding variables both internal and external to experimental subjects. For example, prior musical training might make a difference to how one responds to Mozart music, but the experimenter may not have taken this into consideration in their design. But internal validity can also fail to hold is by virtue of what I have referred to as the problem of conceptual scope (for example, we may refer to the effect as a Mozart effect when it is in fact a Major-key effect). Effectively, when I treat a major-key effect as a Mozart effect, I have misidentified the relevant causal feature of the stimulus. In turn, this means that I will neglect to control for major/minor key as I will regard this as irrelevant, which can result in systematic errors. In both cases, scientists can go wrong in their individuation judgment. What is at stake is not whether there is an effect, but what kind of effect it is. Now, given that those kinds of problems can

[4] In this respect I differ from some advocates of conceptual replication, who have highlighted external validity as a desideratum (E.g., Lynch 1982, 3/4).

occur, we turn to the question of whether "conceptual replication" has an answer. I will now argue that it does not.

To explain this, let me return to the above characterization of conceptual replication, according to which such replication consists in repeating an experiment, using different operationalizations of the same construct. For example, a conceptual replication of an experiment about the Mozart effect might operationalize the concept Mozart effect differently by using a different piece of Mozart music and/or a different measure of spatial reasoning. But there is a major caveat here: If I want to compare the results of two experiments that operationalized the same construct differently, I already have to presuppose that both operationalizations in fact have the same conceptual scope, i.e., that they in fact individuate the same effect. But this would be begging the question, since after all – given the epistemic uncertainty and conceptual openness highlighted above – that's precisely what's at issue. Differently put, experiment 2 might or might not achieve the same result as experiment 1, but the reason for this would be underdetermined by the experimental data. Thus, the problem of conceptual scope prevents us from being able to say whether we have succeeded in our conceptual replication.

Given the uncertainties as to whether one has in fact succeeded in conceptually replicating a given experiment, I am weary of the language of replication here. If anything, I would argue that the method in question should be regarded as a research strategy that is aimed at helping to demarcate and explore the very subject matter under investigation. But as I will argue now, this is perhaps better described as exploration, not as replication.

4. Putting Replication in its Proper Place

The conclusion of the previous paragraphs seems pretty bleak: Direct replication is either extremely narrow in what it can deliver or it runs into the joint problems of confounders and conceptual scope. Conceptual replication, on the other hand, cannot come to the rescue, because it also runs into the

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -228-

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

exact same problems. Should we then throw up our hands and conclude that since ultimately neither direct nor conceptual replication are possible the crisis of replication is much more severe than we previously thought? This would be the wrong conclusion, however. This would only follow if replication was in fact as central to research as it is sometimes taken to be. I claim that it is not. My argument for these claims has three parts. The first part holds that exploring (the possibility of) systematic errors is an important part of the investigative process, which is not well described as replication. Second, if we take seriously this process of exploring and delineating the relevant phenomena, we find that there is indeed a great deal of uncertainty in psychological research, but this, in and of itself, does not necessarily constitute a crisis. Lastly, while it is fair to say that there is a crisis of confidence in current psychology, it is not well described as a replication crisis.

Let me begin with the first point. I have argued that direct replication (even where it is successful) is of limited value, because it can at most rule out random error, but completely fails to be able to address systematic error. But if we appreciate (as I have argued we should) that direct replication inevitably involves individuation judgments, it is obvious that there is always a danger of systematic error, because I have to assume that all confounding variables have been controlled for. One important class of confounders follows from what I have referred to as the problem of conceptual scope, i.e., the difficulty of correctly describing both the independent variable responsible for a given effect and the dependent variable.[5] Epistemically productive experimental work, I claim, therefore needs to focus on systematic errors, specifically those brought about by unstated auxiliary assumptions.

Indeed, if we look at the story of the Mozart effect, we find that this is exactly what happened. This example also nicely illustrates my claim about the conceptual openness and epistemic uncertainty

[5] My focus here has been mainly on the former. But of course the problem of conceptual scope concerns both.

in many areas of experimental psychology. The Mozart effect was first posited by Rauscher and colleagues (Rauscher et a. 1993). It can now be regarded as largely debunked. While it is true that several people tried (and failed) to replicate the effect (e.g., Newman et al. 1995; Steele 1997), it is important to look at the details here. It is not the case that the effect was simply abandoned for lack of replicability. Rather, when we look at the back and forth between Rauscher and her critics, we find that the discussion turned on the choices and interpretations of independent and dependent variables. In this vein, Newman et al (1995) and Steele (1997) used different dependent variables, prompting Rauscher to argue that her effect was more narrowly confined to the kind of spatial reasoning measured by the Stanford-Binet. I suggest that we interpret this case as one where Rauscher was forced to confront (and retract) an unstated auxiliary assumption of her initial study, namely that the spatial reasoning subtest of the Stanford-Binet (which she had used as her dependent variable), was representative of spatial reasoning more generally. Likewise, her choice of the Mozart's Sonata for Two Pianos in D-major as the independent variable was put under considerable pressure by critics, who suggested that the relevant feature of the independent variable was not that it was a piece by Mozart, but that it was up-beat and put subjects in a good mood (Chabris 1999). My point here is that the debates surrounding the Mozart effect are best described as conceptual work, exploring consequences of possible errors that might have arisen from the problem of conceptual scope. At issue, I claim, was not primarily whether Rauscher really found an effect, but rather what was the scope of the effect.

I argue that this is a typical case. Rather than, or in addition to, attempting to conduct direct replications of previous experiments, researchers critically probed some hidden assumptions built into the design and interpretation of the initial experiment. My point here is both descriptive and normative. Thus, I argue that this is a productive way to proceed. However, I claim that it is not well described as replication, let alone conceptual replication. Rather, what we see here is a case in which scientists explore the empirical contours of a purported effect in the face of a high degree of epistemic

uncertainty and conceptual openness, and this is precisely why the case is not well described as employing conceptual replication. The reason for this is quite simple: For a conceptual replication to occur, one needs to already be in the possession of some well-formed concepts, such that they can be operationalized in different ways. It also presupposes that in general the domain is well-understood, such that operationalizations can be implemented and confounding variables can be controlled. But this completely misses the point that researchers often investigate effects precisely because they don't have a good understanding (and hence concept) of what it is.

Therefore I argue that while direct replication can only contribute a very small part to the research process, conceptual replication cannot make up for the shortcomings of direct replication. Instead, productive research should (and frequently does) proceed by exploring, and experimentally testing, hypotheses about possible systematic errors in experiment. Such research, I suggest, can contribute to conceptual development by helping to explore and fine-tune the shape and scope of proposed or existing concepts. The fact that this is riddled with problems does not in and of itself constitute a crisis, let alone a replication crisis.

5. Conclusion

The upshot of the above is that when we talk about the importance of replication, we need to be clear on what we mean by replication and why it is so important, precisely.

In this paper I have argued that if by replication we mean either "direct" or "conceptual" replication, we need to first of all be clear that direct replications are not non-conceptual. I then turned to some alleged epistemic merits of direct replication, for example that they can establish the existence of effects or the reliability of procedures that detect effects. I argued that insofar as such replications involve concepts, they run (among other things) into the problem of conceptual scope, i.e., the difficulty of determining, on the basis of independent and dependent variables of experiments what precisely is

the scope of the effect one is trying to replicate. I highlighted that this is a real and pernicious problem in experimental research in psychology, due to the high degree of epistemic uncertainty and conceptual openness of many fields of research.

While my emphasis of the conceptual nature of replication may suggest that I would be more favorably inclined toward conceptual replication, I have argued that conceptual replication runs into the same problems, and for similar reasons: The very judgement that one has successfully performed a conceptual replication of a previous experiment presupposes what is ultimately the aim of the research, namely to arrive at a robust understanding of the relevant area of research. This, I argue that since conceptual replication presupposes a relatively good grasp of the relevant concepts, it is begging the question, and I suggested instead that researchers (should) engage in a process of specifically investigating possible systematic errors in original studies as a means to develop the relevant concepts. This process is not best described as one of replication, however. Summing up, then, I conclude that in general, replications are less useful and important than is widely assumed – at least in the kind of psychological research I have focused on in this paper.

Now, in conclusion let me return to the notion of a crisis in psychology as it is currently discussed in the literature. Obviously, I do not mean to deny that there is a crisis of confidence in (social) psychology (Earp & Trafimov 2015) as well as in other areas of study. However, based on the analysis provided in this paper, I argue that this crisis is not well described as a crisis of replication. Rather, it seems to be to a large degree a crisis that turns on questionable research practices with regard to the use of statistical methods in psychology (see Gelman & Loken 2014). While acknowledging the valuable philosophical and scientific work that is being done in this area, I suggest that a broader focus on the notion of replication provides us with a deeper appreciation of the conceptual dynamics characteristic of experimental practice.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -232-

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

REFERENCES

Campbell, D. T., and Stanley, J. C. (1966), Experimental and Quasi-Experimental Designs for Research (Chicago: Rand McNally).

Chabris, C. (1999): Prelude or requiem for the 'Mozart Effect?' "Scientific Correspondence", Nature, 4000, 826.

Collins, H. (1985). Changing order. Replication and induction in scientific practice. Chicago and London: The University of Chicago Press.

Earp, Brian & Trafimow, David (2015): "Replication, falsification, and the crisis of confidence in social psychology." Front. Psychol, 19 May 2015 | https://doi.org/10.3389/fpsyg.2015.00621

Feest, U., 2016, "The Experimenters' Regress Reconsidered: Tacit Knowledge, Skepticism, and the Dynamics of Knowledge Generation". Studies in History and Philosophy of Science, Part A 58 34-45.

Feest, U. & Steinle, F., 2016, "Experiment." In P. Humphreys (Ed.): Oxford Handbook of Philosophy of Science. Oxford University Press, 274–295.

Fiedler, K.; Kutzner, F. & Krueger, J. (2012): „The Long Way from alpha-error control to validity proper: Problems with a short-sighted false-positive debate." Perspectives on Psychological Science 7(6), 661-669

Gelman, Andrew & Loken, Eric (2014): The Statistical Crisis in Science. Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up. American Scientist 102 (6) 460-464. DOI: 10.1511/2014.111.460

Goodman, Nelson (1983/1955): Fact. Fiction and Forecast. Harvard University Press; 4 Revised edition edition

Guala, F. (2012), "Philosophy of Experimental Economics." In U. Mäki (ed.), Handbook of the philosophy of science. Vol. 13: Philosophy of Economics (Boston: Elsevier/Academic Press), 597–640

LeBel, E.P.; Berger, D., Campbell, L.; Loving, T. (2017): "Falisifiability is not Optional." Journal of Personality and Social Psychology (forthcoming)

Lynch, J. (1982): "On the External Validity of Experiments in Consumer Research. Journal of Consumer Research 9, 225-239. (December)

Lynch, J.; Bradlow, E.; Huber, J.; Lehmann, D. (2015): "Reflections on the replication corner: In praise of conceptual replication." IJRM ???

Mayo, Deborah (1996): Error and the Growth of Experimental Knowledge. University of Chicago Press.

Newman, J., Rosenbach, J., Burns, K.; Latimer, B., Matocha, H., Vogt, E. (1995: An experimental test of the 'Mozart Effect': Does listening to Mozart improve spatial ability? Perceptual and Motor Skills, 81, 1379-1387.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349

Pashler, Harold & Harris, Christine (2012): "Is the Replication Crisis Overblown?" Perspectives on Psychological Science 7(6), 531-536.

Rauscher, F., Shaw, G.; Ky, K. (1993). Music and spatial task performance. Nature ,365, 611.

Romero, Felipe (2017): "Novelty vs. Replicability. Virtues and Vices in the Reward System of Science." Philosophy of Science.

Shavit, Ayelet & Ellison, Aaron (eds.) (2017): Stepping in the Same River Twice. Replication in Biological Research. Yale University Press

Soler, Lena (2011): "Tacit Elements of Experimental Practices: analytical tools and epistemological consequences." European Journal for Philosophy of Science 1, 393-433.

Steele, K., (2000). Arousal and mood factors in the 'Mozart effect'. Perceptual and Motor Skills, 91, 188-190.

Zwaan, Rolf (2013): "How Valid are our Replication Attempts?"

https://rolfzwaan.blogspot.de/2013/06/how-valid-are-our-replication-attempts.html

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -234-

Author: Paul L. Franco, UW-Seattle, Department of Philosophy
Contact: pfranco@uw.edu

Title: Speech Act Theory and the Multiple Aims of Science

Abstract: I draw upon speech act theory to understand the speech acts appropriate to the multiple aims of scientific practice and the role of nonepistemic values in evaluating speech acts made relative to those aims. First, I look at work that distinguishes explaining from describing within scientific practices. I then argue speech act theory provides a framework to make sense of how explaining, describing, and other acts have different felicity conditions. Finally, I argue that if explaining aims to convey understanding to particular audiences rather than describe literally across all contexts, then evaluating explanatory acts directed to the public or policymakers involves asking nonepistemic questions.

(Accepted with minor revisions to the PSA 2018 proceedings issue of Philosophy of Science | Revisions not yet made; final version due January 2019)

1. Introduction

Hasok Chang "[complains] about…our [i.e., philosophers of science] habit of focusing on descriptive statements that are either products or presuppositions of scientific work, and our commitment to solving problems by investigating the logical relationships between these statements" (2014, 67–8). He argues philosophers of science should adopt "a change of focus from propositions to actions" (67). Chang suggests, "When we do pay attention to words, it would be better to remember to think of 'how to do things with words', to recall J. L. Austin's (1962) famous phrase" (68).

In this paper, I take Chang's suggestion and argue that attending to Austin's account of the things we do with words can help us understand the multiple goals of scientific practices, the speech acts appropriate to those goals, and the roles of nonepistemic values in evaluating speech acts made relative to those aims. In §2, I give an overview of a few philosophers of science working on explanation who have shifted focus from propositions to explaining.[1] I also briefly relate this work to a few themes in speech act theory. In §3, I give more details of Austin's framework to highlight ways of evaluating speech acts beyond truth and falsity. In §4, I explore the multiple goals of scientific practice, especially goals related to conveying understanding to the general public and policymakers, and the speech acts appropriate to those goals.

2. The things scientists do with words

2.1 Explaining

Consider some recent and not-so-recent work on scientific explanation. Andrea Woody's defense of a functional perspective on explanation aims to motivate "a shift in focus away from explanations, as achievements, toward explaining, as a coordinated activity of communities" (2015, 80). In a similar spirit, Angela Potochnik argues that when looking at explanation, "sidelining the communicative purposes to which explanations are put is a mistake" (2016, 724). She emphasizes that explaining is a communicative act involving a speaker and audience made against a background that shapes the explanations offered. In so

[1] I make no claims Chang influenced the work I canvas.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -236-

arguing, Potochnik deliberately recalls Peter Achinstein's claim, "Explaining is an illocutionary act," i.e., a speech act uttered by a speaker with a certain force and for a certain point (1977, 1).

These accounts share in common an emphasis on the importance of the aims of the speaker and audience, and thus the context of utterance in evaluating, to borrow terminology from Austin, the felicity conditions of explanatory speech acts. In particular, we might focus on the aims of the speaker and their audience in requesting and giving explanations, the time and location of an explaining speech act, and, following Woody, "what role(s) [explanations] might play in practice" (2015, 81). In focusing on the explaining act rather than the supposedly stable propositional content of an act of explanation, our attention is drawn to dimensions of evaluation beyond truth and falsity.

On this last point, Nancy Cartwright argues that the functions of a scientific theory to "tell us…what is true in nature, and how we are to explain it…are entirely different functions" (1980, 159). *Ceteris paribus* laws used in scientific theories are literally false, but still do explanatory work. One way to understand Cartwright's claim is that the speech act of describing the world truly and the speech act of explaining come apart from one another. In coming apart from one another and fulfilling different aims within scientific practice, descriptive and explanatory speech acts have different felicity conditions. For example, Potochnik (2016) examines the ways in which explaining increases understanding. But, Potochnik argues, what gets explained depends on a speaker's and audience's interests, and an explaining act's success in generating understanding depends on the cognitive resources of the audience. As such, to evaluate any given communicative act of explaining requires attending to the epistemic and nonepistemic interests of speakers and audiences that form the background against which explanations are offered. This means evaluating explanatory speech acts solely in terms of truth or falsity is inapt.

2.2 Multiple aims and the true/false fetish

I do not think this focus on acts and away from the truth or falsity of descriptive statements is unique to philosophers of science interested in explanation. We see a similar shift in work on the so-called aims approach to values in science (e.g., Elliott and McKaughan 2014;

Intemann 2015). The aims approach shares in common with work on explaining a recognition that scientific practice aims at more than describing the world truly or falsely. Further, if some of those aims include things like making timely policy recommendations for decision makers or increasing public understanding of science, there is a role for nonepistemic values in parts of scientific practice. As Kevin Elliott and Daniel McKaughan put this point, "representations can be evaluated not only on the basis of the relations that they bear to the world but also in connection with the various uses to which they are put" (2014, 3).

Why look to speech act theory to flesh out this picture about the multiple aims of scientific practice and their relationship to nonepistemic values? In part because speech act theory makes sense of the different uses to which one and the same sentence might be put depending on the aims of the speaker and audience and the context of utterance. In doing so, I think Austin is right that we can "play Old Harry with two fetishes…(1) the true/false fetish, (2) the value/fact fetish" (1962, 150). Austin was mainly content to play Old Harry with these fetishes to free philosophers from the grip of the so-called descriptive fallacy: the view "that the sole business, the sole interesting business, of any utterance…is to be true or at least false" (1970, 233). But I also think that in combating the descriptive fallacy and the true/false and fact/value fetishes, speech act theory motivates a constructive shift from the truth or falsity of descriptive statements to the things we do with words.

Take Austin's claim that evaluating apparently descriptive speech acts like "'France is hexagonal,'" involves nonepistemic questions about who is uttering the statement, in what context, and with what "intents and purposes" (1962, 142). Rather than concluding the sentence is false and leaving it at that, Austin points out the different speech acts one can use such a sentence to perform, e.g., stating or interpreting or estimating. In determining the use the sentence is put to—with the help of context and by inquiring after the interests of the speaker and their audience—we might realize, irrespective of the sentence's literal truth or falsity, "It is good enough for a top-ranking general, perhaps, but not for a geographer" (142). In other words, it serves the aims of the general, which, unlike the aims of the geographer, do not necessarily require a descriptively literal account of France's shape. The statement might not aim to assert or describe literally, but do something else entirely. As such,

evaluating it along the lines of truth or falsity will miss something important about the aims of a speaker in uttering it.

To expand on this picture, I turn to explicating Austin's speech act theory.

3. Austin's speech act theory

3.1 Performatives and constatives

Austin first drew our attention to the things we do with words by discussing performative utterances. Austin says of these, "if a person makes an utterance of this sort we should say that he is *doing* something rather than merely *saying* something" (1970, 235). Imagine a speaker utters 'I promise to return my referee report in two weeks' during the peer review process. In making this speech act, Austin claims the speaker does not describe an internal act she has concurrent to her utterance. Instead, in making that utterance, the speaker just is performing the act of promising thereby committing herself to actions related to the timely review of papers.

While promising has no special connection to truth and falsity, it still must meet what Austin calls felicity conditions to be happy or unhappy. In order to promise to return their referee report in two weeks successfully, the speaker must meet the sincerity condition of forming an intention to do so, even if they are not describing "some inward spiritual act of promising" (236). The speaker must also be in a position to follow through on their intention. Thus, there is unhappiness in the speech act if the speaker promises knowing full well other commitments will prevent her from returning the report in two weeks. The speaker must also have the authority to make a promise; unless authorized, an editor cannot promise on behalf of a reviewer. There should also exist a convention for making a promise in peer review contexts. Such conventions might allow the speaker to promise without uttering, 'I promise,' e.g., by accepting a request that reads, 'In accepting this review assignment you commit to returning the referee report within such-and-such a time.'

Austin first contrasts performatives with constatives, e.g., descriptive statements or assertions that aim to state something truly or falsely about the world, but which do not seem to perform an action. However, Austin claims describing or asserting is as much an action as promising, even if the felicity conditions for asserting are more closely connected to truth

or falsity. Consider an editor saying of a reviewer, 'They review quickly, and I expect that they will return their review within two weeks.' In saying this, the editor commits herself to providing evidence for her description of the reviewer as quick, and perhaps justifying her expectation that the reviewer's past behavior provides good evidence for future behavior. As Robert Brandom puts this point, "In asserting a claim one not only authorizes further assertions, but commits oneself to vindicate the original claim, showing that one is entitled to make it" (1983, 641). That is, the utterer must be in a position of authority—here in an epistemic sense—with regards to the claim and be ready to perform further speech acts if so prompted. Other felicity conditions of assertions or descriptions include a sincerity condition: an editor uttering our example sentence should believe what they say. Finally, the context of an assertion also shapes its felicity conditions: an editor should utter the sentence in the appropriate circumstances, e.g., as a response to a worry about the speed of the review process. Should these conditions not be met, the speech act might be unhappy even if true.

3.2 Locution and illocution

Austin develops speech act theory to capture the similarities between performatives and constatives. Speech acts like promising and describing have three dimensions: the locutionary content, which is the conventional sense and reference of the uttered sentence; the illocutionary force, which is the use the utterance is put to; and the perlocutionary effects, which are intended and unintended "effects upon the feelings, thoughts, or actions of the audience, or of the speaker, or of other persons" (1962, 101).

Austin's points about the illocutionary dimension of a speech act most clearly capture how one and the same representation might be put to different uses depending on our goals, and how different uses have different felicity conditions despite sharing locutionary content. Consider the sentence, 'This product contains chemicals known to the state of California to cause cancer.' The locutionary content would just consist in the proposition expressed by the sentence as determined by the conventional sense and reference of the words. This content can be common to different illocutionary acts. Someone uttering the sentence could be describing a product, issuing a warning, or explaining why they do not use this particular product but another. Uttering the sentence with the force of a description, the force of a

warning, and the force of an explanation will have similar felicity conditions related to truth and falsity. Namely, the locutionary content should be true or approximately true for an utterance to count as a good description, a good warning, or a good explanation.

However, a warning might be infelicitous in ways a description might not. For example, warnings might be issued only in the case in which some pre-determined level of significant risk at a certain level of exposure is met. In cases where such levels are not met, issuing a warning might be infelicitous. Consider also that uttering such a sentence with the force of an explanation might be called for only if, e.g., someone is prompted to justify their choice of a product that does not contain cancer-causing chemicals over a more easily available and cheaper product that does contain those chemicals. In these last two cases, nonepistemic reasons related to risk, cost-effectiveness, and so on can enter into the evaluation of the happiness of a warning or explanation.[2]

Austin thinks attending to these points combats a form of abstraction that distorts our thinking about the felicity conditions of descriptive statements. He thinks that when examining statements, "we abstract from the illocutionary…aspects of the speech act, and we concentrate on the locutionary" (1962, 144–5). In so doing, "we use an over-simplified notion of correspondence with the facts—over-simplified because essentially it brings in the illocutionary aspect" (145). Such an approach focuses on "the ideal of what would be right to say in all circumstances, for any purpose, to any audience, &c." (145). But, as Austin claims, questions concerning correspondence with the facts brings with it the illocutionary aspect since truth or falsity does not attach to sentences or locutionary content. Instead, truth or falsity is related to particular things speakers do with sentences. Descriptions might be, strictly speaking, true or false, but not recommendations or explanations. In order to know, then, if evaluating a speech act along the true-false dimension is apt, we need to know the illocutionary force of that act. But to know the illocutionary force of the act requires we attend to context, including the aims of both speaker and audience, time and place of utterance, and conventions governing the specific speech situation. In this way, Austin

[2] Any speech act will also have perlocutionary effects, and we might follow Heather Douglas (2009) and Paul Franco (2017) in focusing on the nonepistemic consequences of making false descriptions, giving bad warnings, or explaining unclearly.

argues context and aims are central to determining the illocutionary force of a speech act, and hence to evaluating its felicity or infelicity.

4. Aims-approaches and speech act theory

4.1 Explaining and understanding

Scientific practice might seem to deal in paradigmatically constative speech acts, e.g., descriptions. Such speech acts are, to varying degrees, evaluable along dimensions of truth or falsity in ways we might question the relevance of speech act theory to philosophy of science. That is, we might say that scientific practice just is a case in which abstracting away from the illocutionary force of an utterance to focus on locutionary content is appropriate. For example, Austin says that "perhaps with mathematical formulas in physics books…we approximate in real life to finding" speech acts where focusing on the locutionary content is appropriate (1962, 145). If scientific practice aims at timeless truths holding across all contexts independent of the sorts of aims and interests of speakers and audiences necessary to evaluating the felicity or infelicity of speech acts, then it seems speech act theory is irrelevant to philosophy of science.

Yet, as Austin points out, "When a constative is confronted with facts, we in fact appraise it in ways involving the employment of a vast array of terms which overlap with those that we use in the appraisal of performatives. In real life, as opposed to the simple situations envisaged in logical theory, one cannot always answer in a simple manner whether it is true or false" (141–2). Consider again 'France is hexagonal.' Austin asks, "How can one answer…whether it is true or false that France is hexagonal? It is just rough, and that is the right and final answer to the question of the relation of 'France is hexagonal' to France. It is a rough description; it is not a true or false one" (142). Though rough, it is still open to evaluation. We can ask if it is in accord with conventions governing estimations and if this estimation serves the purposes and interests of the speaker and their audience at the time of utterance. 'France is hexagonal' can count as felicitous even if rough and not literally true because it aims at something other than truth.

Austin claims that many of our apparently constative speech acts are evaluable along similar dimensions given that they also confront facts in similarly rough ways. McKaughan

makes a related point about scientific speech acts. He argues that certain speech acts central to scientific practice like "conjecturing, hypothesizing, guessing and the like often play a role in scientific discourse that serves neither to assert that an hypothesis is true nor to express such a belief" (2012, 89). Moreover, as mentioned in §2, the picture of scientific practice as concerned solely with the truth is challenged, among other places, in work on explanation, and also in values in science. For example, when looking at the role particular acts or patterns of explaining play in scientific discourse we might focus not on the locutionary content of an explanatory speech act, but on the ways "explanatory discourse…functions to sculpt and subsequently perpetuate communal norms of intelligibility" (Woody 2015, 81). In focusing on this aspect of explaining, we might find, for example, that "the ideal gas law's role in practice is not essentially descriptive, but rather prescriptive; by providing selective attention to, and simplified treatment of, certain gas properties (and their relations) and ignoring other aspects of actual gas phenomena, the ideal gas law effectively instructs chemists in how to think about gases as they are characterized within chemistry" (82). In other words, the ideal gas law, in practice, does not have the force of a descriptive speech act, but lays down a rule of sorts guiding the investigation of gases.[3] The success of acts of explaining from this perspective will have less to do with accurately describing actual gases, but the way they facilitate, say, the education of new scientists or increase understanding of related phenomena, e.g., "by laying foundation for the concept of 'temperature'" beyond "the subjective, inherently comparative quality of human perception" (82). An act of explaining that fails to achieve pedagogical aims or fails to increase understanding of related phenomena might be infelicitous even if the locutionary content of that act confronts the facts in the right way to count as approximately true.

On this point about the ways explanations might increase understanding without describing, Potochnik claims "that what best facilitates understanding is not determined solely by the relationship between a representation and the world" (2015, 74). An idealized explanation like the ideal gas law is not defective because it fails to fully describe all the

[3] About universal generalizations Austin writes, "many have claimed, with much justice, that utterances such as those beginning 'All…' are prescriptive definitions or advice to adopt a rule" (1962, 143). Austin does not fully endorse this suggestion.

possible causal factors at play in the behavior of actual gases. Though literally false, an idealization might be successful insofar as it "secure[s] computational tractability" or successfully isolates "all but the most significant causal influences on a phenomenon" (71). In so doing, we increase our understanding by facilitating "successful mastery, in some sense, of the target of understanding" or "by revealing patterns and enabling insights that would otherwise be inaccessible" (72). Indeed, pointing out all the ways in which the ideal gas law fails to hold for actual gases or is literally false as a description might hinder the use of explanations in scientific discourse to provide "shared exemplars that function as norms of intelligibility" (Woody 2015, 84).

In a related vein, Potochnik argues, "Because understanding is a cognitive state, its achievement depends in part on the characteristics of those who seek to understand," including both the speaker and the audience (2015, 74). In evaluating an act of explaining, we should look at how the speaker's interest has shaped the focus of their explanation and also how the explanation increases an audience's understanding, where this involves considering the audience's interests in seeking an explanation. An explanation that fails to be relevant to the audience or fails to increase their understanding or guide their thinking about related phenomena, but that nonetheless has locutionary content that is approximately true, might count as infelicitous.

4.2 Values and science

On the views of explaining canvassed, the aims of generating literally true descriptions of the world come apart from, say, explaining and understanding the most important causal factors at play for a given phenomenon. Now, as the aims approach to the proper role for nonepistemic values in scientific practice emphasizes, explaining and describing do not exhaust the goals of scientific practice. The aims approach focuses on the ways "scientific decision-making, including methodological choices, selection of data, and choice of theories or models, are...a function of the aims that constitute the research context" (Intemann 2015, 218). Given that the research context includes social, political, and moral considerations, the aims of science can just as well be understood in nonepistemic ways as it can be understood in epistemic ways.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -244-

Consider, for example, the American Geophysical Union's position statement on human-induced climate change. At the end of their statement, they claim, "The community of scientists has responsibilities to improve overall understanding of climate change and its impacts. Improvements will come from pursuing the research needed to understand climate change, working with stakeholders to identify relevant information, and conveying understanding clearly and accurately, both to decision makers and to the general public" (American Geophysical Union 2013). Here, I focus on the claim that scientists have responsibilities to improve the understanding of policymakers and the general public, and drawing upon the aforementioned work on explaining, think about how adopting this aim shapes the felicity conditions of explanatory speech acts directed at the audiences mentioned.

Notice that the position statement distinguishes the research necessary to understand climate change from conveying that understanding to policymakers and the general public. The sense in which these different activities come apart from one another and have different success conditions can be made sense of, in part, by focusing on the audience to whom scientists are speaking. We saw that for Potochnik (2016) understanding is a cognitive state that depends on the abilities and interests of those who are explaining and those to whom explanations are directed. In communicating to policymakers and the general public, scientists should consider the interests of the speaker in asking for an explanation as well as their level of knowledge regarding the phenomenon in question, in this case, climate change. In so doing, scientists might find that a description that aims to describe climate change in all its complexity might not serve these aims well. Instead, scientists might aim for an explanation that, though omitting descriptive complexity, draws upon models that represent those causal factors related to the audience's interests in a way that is cognitively accessible and helps guide the public in thinking more generally about climate change.

On this point, the American Geophysical Union's position statement maintains scientists ought to enlist the help of stakeholders in identifying potentially relevant information to their research. This is a point Intemann makes in developing the aims approach. She says of climate science, "[T]he aim is not only to produce accurate beliefs about the atmosphere, but to do so in a way that allows us to generate useful predictions for protecting a variety of social, economic and environmental goods that we care about" (2015,

219). In the view of the American Geophysical Union, in order to do this well, scientists ought to consult with relevant stakeholders and policymakers regarding what they value. Thus, for example, if stakeholders and policymakers communicate worries about extreme weather events and "how to adapt to 'worst case scenarios,' then models able to capture extreme weather events should be preferred" to those models that "anticipate slow gradual changes" (Intemann 2015, 220). Notice that in making such a decision, the grounds for choosing models able to represent aspects of climate change relevant to stakeholders' interests are nonepistemic rather than epistemic, e.g., generating predictions useful for protecting goods the general public cares about. Insofar as the representations or explanations generated do not meet these goals because they are unrelated to stakeholders' interests, the attendant speech acts might very well be infelicitous even if they describe some related phenomenon more or less accurately.

Both points about pitching explanations at a level that is cognitively accessible and choosing models for representing climate change phenomena in ways sensitive to stakeholders' interests illustrate a point Austin makes about the importance of uptake to successfully performing a speech act. Austin claims, "Unless a certain effect is achieved, the illocutionary act will not have been happily, successfully performed….I cannot be said to have warned an audience unless it hears what I say and takes what I say in a certain sense….Generally the effect amounts to bringing about the understanding of the meaning and force of the locution" (1962, 116). In aiming to convey understanding through explaining relevant aspects of climate change to decision makers and the general public, a speaker should consider the interests, background knowledge, and cognitive resources of their audience. Insofar as scientists fail to do so in explaining to the general public, even if the locutionary content that comprises their speech act approximates truth, they will not secure uptake in the sense of generating understanding in their audience. As such, their speech act will be infelicitous.

Of course, a scientist's explaining something to their audience will also be infelicitous if it is based on inaccurate information or extrapolates from what is known to their audience's interests in unjustified ways. However, this does not mean that if scientists aim to convey understanding to the public they should stick solely to descriptive claims. As

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -246-

Elliott emphasizes in discussing how scientists should best communicate uncertainty to the public, "It does little good to expect scientists to provide unbiased information to the public if their pronouncements are completely misinterpreted or misused by those who receive them" (2017, 89). Similarly, "members of the public might not be able to 'connect the dots'" between scientists' descriptive speech acts and the ways those are relevant to their interests; insofar as scientists do not explain with the aims of conveying understanding—which as Potochnik argues, comes apart from describing the world truly in all its complexity—the public "would be left wondering what [the descriptions] might mean" (88). Thus, if scientists are to meet responsibilities the American Geophysical Union claims they have with regard to conveying understanding to the general public, those scientists should communicate using speech acts best able to secure uptake in the general public. This involves considering the interests and cognitive resources of the general public in ways that shape the felicity conditions of the speech acts beyond truth and falsity.

5. Conclusion

I argued speech act theory can tie together a few threads in recent work on explaining and values in science that share in common a shift in focus from descriptive propositions to things scientists do with words. Some of those things, like explaining, also seem the sorts of speech acts appropriate for fulfilling aims scientists have other than describing the world literally, like conveying understanding to the public and policymakers. Insofar as successfully fulfilling these aims involves explaining, and insofar as acts of explaining that secure uptake require attention to the nonepistemic interests and cognitive resources of speaker and audience, our attention is drawn towards ways explanatory speech acts can be happy or unhappy beyond describing truly or falsely. Future work will aim to delineate these felicity conditions in greater detail with an eye towards revealing further nonepistemic dimensions of evaluation.

References

Achinstein, Peter. 1977. "What is an Explanation?" *American Philosophical Quarterly* 14(1):1–15.

American Geophysical Union. 2013. "Human-Induced Climate Change Requires Urgent Action." https://sciencepolicy.agu.org/files/2013/07/AGU-Climate-Change-Position-Statement_August-2013.pdf

Austin, J.L. 1962. *How to Do Things With Words*. Ed. J.O. Urmson. Oxford: Oxford University Press.

----. 1970. "Performative Utterances." *Philosophical Papers*, 2nd edition. Eds. J.O. Urmson and G.J. Warnock. Oxford: Oxford University Press: 233–252.

Brandom, Robert. 1983. "Asserting", *Nous* 17(4):637–650.

Cartwright, Nancy. 1980. "The Truth Doesn't Explain Much." *American Philosophical Quarterly* 17(2):159–163.

Chang, Hasok. 2014. "Epistemic Activities and Systems of Practice: Units of Analysis in Philosophy of Science After the Practice Turn." *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science*, eds. Léna Soler, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost. New York: Routledge: 67–79.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, Kevin. 2017. *A Tapestry of Values*. New York: Oxford University Press.

Elliott, Kevin C. and Daniel J. McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81(1):1–21

Franco, Paul L. 2017. "Assertion, Nonepistemic Values, and Scientific Practice." *Philosophy of Science* 84(1):160–180.

Intemann, Kristen. "Distinguishing Between Legitimate and Illegitimate Values in Climate Modeling." *European Journal of the Philosophy of Science* 5:217–232.

McKaughan, Daniel J. 2012. "Speech acts, attitudes, and scientific practice: Can Searle handle 'Assuming for the sake of Hypothesis'?" *Pragmatics and Cognition* 20:1:88–106.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -248-

Potochnik, Angela. 2015. "The Diverse Aims of Science." *Studies in History and Philosophy of Science Part A* 53:71–80

----. 2016. "Scientific Explanation: Putting Communication First." *Philosophy of Science*, 83:721–732.

Woody, Andrea. 2015. "Re-orienting discussions of scientific explanation: A functional perspective." *Studies in History and Philosophy of Science Part A* 52:79–87.

Universality Reduced

Alexander Franklin*†

October 2018

Forthcoming in *Philosophy of Science: Proceedings of the PSA 2018*

Abstract

The universality of critical phenomena is best explained by appeal to the Renormalisation Group (RG). Batterman and Morrison, among others, have claimed that this explanation is irreducible. I argue that the RG account is reducible, but that the higher-level explanation ought not to be eliminated. I demonstrate that the key assumption on which the explanation relies – the scale invariance of critical systems – can be explained in lower-level terms; however, we should not replace the RG explanation with a bottom-up account, rather we should acknowledge that the explanation appeals to dependencies which may be traced down to lower levels.

1 Introduction

While universality is best explained with reference to the Renormalisation Group (RG), that explanation is nonetheless reducible. The argument in defence of this claim is of philosophical interest for two reasons: first, the RG explanation of universality has been touted by Batterman (2000, 2017) and

*alexander.a.franklin@kcl.ac.uk

†I am grateful to Eleanor Knox, and to the audience of the IMPS 2018 conference in Salzburg for helpful comments. This work was supported by the London Arts and Humanities Partnership.

Morrison (2012, 2014) as a significant impediment to reduction. Second, universality is a paradigm instance of multiple realisability (MR) in the philosophy of physics; as such it is regarded as irreducible by those who accept the multiple realisability argument against reduction. My account charts a middle course: I deny claims that RG explanations are irreducible, and I deny that universality is *best* explained from the bottom up.

The view of reduction advocated here is non-eliminativist; the best explanations are often higher-level explanations: such explanations are more parsimonious, more robust, and have broader applicability than lower-level explanations. In general, such higher-level explanations ought not to be replaced by lower-level explanations, rather the parts of theories on which such explanations rely may be understood in lower-level terms; reducible explanations satisfy the following two conditions: (a) each higher-level explanatory dependency is explained by or derived from a lower-level dependency, and (b) the abstractions involved in constructing the higher-level explanations are justified from the bottom up.[1]

In §2 I outline the RG explanation of universality. Although my reductive claims may generalise, I focus exclusively on the field-theoretic approach to the RG.[2] I claim that this explanation follows a general formula for explaining multiply realised phenomena. §3 considers the arguments of Batterman and Morrison, and analyses their force against any putative reduction.

In §4 I note that the RG explanation is a higher-level explanation. As it is less contentious that the common features of each universality class are reducible, I simply assume that that's the case in this paper. The nub of the debate rests on the RG: I show that the RG arguments rely on the assumption of scale invariance and the abstractions engendered by that assumption. I argue that the applicability of this assumption may be explained from the bottom up. Thus, I claim, that my reduction satisfies (a) and (b) above.

[1] While I expect the claims in this paper to be compatible with many different accounts of explanation, they are most straightforwardly cashed out on an interventionist approach – see Woodward (2003).

[2] See Franklin (2018) and Mainwood (2006) for arguments that only this approach provides an adequate explanation of universality.

2 The RG Explanation of Universality

'Universality' refers to the phenomenon whereby diverse systems exhibit similar scaling behaviour on the approach to a continuous phase transition. Continuous phase transitions occur at the critical temperature, a point beyond which systems no longer undergo first-order phase transitions.[3] The approach to this phase transition can be very well described by power laws of the form $a_i(t) \propto t^{\alpha}$ where t is proportional to the temperature deviation from the critical temperature and α is the critical exponent – a fixed number which leads to a characteristic curve on temperature-density plots.[4]

Different physical systems can be categorised into universality classes: members of the same class have identical critical behaviour – the same set of critical exponents $\{\alpha, \beta, ...\}$ for several power laws – while their behaviour away from the critical point and microscopic organisation may be radically different. For example, fluids and magnets are in the same universality class despite otherwise having totally different chemical and physical properties.

Each physical system which exhibits critical phenomena may be described at the critical point by the same mathematical object – the Landau-Ginzburg-Wilson (LGW) Hamiltonian. That Hamiltonian will include the features – the symmetry and dimensionality – which sort these systems into their universality classes. The RG argument demonstrates that the LGW Hamiltonian applies to a wide range of systems at the critical point by showing that any additional operators which may be appended to that Hamiltonian will fall away on approach to criticality, where only the central LGW operators will remain. The following steps are essential to the explanation thus on offer:[5]

1. Define the effective Hamiltonian for your system of interest:
   (i) Specify the order parameter with symmetry and dimensionality.
   (ii) Specify the central operators of the LGW Hamiltonian.

[3] Note that not all continuous phase transitions are associated with first-order phase transitions in this way.

[4] E.g. the specific heat (in zero magnetic field) c scales as $c \sim (t^{-\alpha})/\alpha$ as $t \to 0$ where $t = \frac{T - T_c}{T_c}$.

[5] To see a full account of the physics of universality and details of the RG see Binney et al. (1992) and Fisher (1998); the philosophical aspects of such an explanation are discussed in detail in Batterman (2016) and Franklin (2018).

(iii) Specify operators in addition to the terms in the LGW Hamiltonian.

2. Apply the RG transformations to that Hamiltonian.
3. Examine the flow towards fixed points in the critical region and note that some operators are irrelevant to the critical behaviour.
4. Thus divide the set of operators into subsets: 'relevant', 'irrelevant' and 'marginally relevant'.
5. Repeat for other systems of interest.

In order to explain universality we must identify commonalities between the different systems in the same universality class – 1(i) and 1(ii) above – and show that such commonalities are sufficient for the common behaviour – 2-4 above. Although 1(iii) can't, in general, be done explicitly, the explanation only depends on the RG demonstration that all distinguishing features are irrelevant – it's not necessary to say exactly which those distinguishing features are. As discussed below, the infinities which are central to some of the anti-reductionist arguments feature in steps 3 and 4.

Overall the explanation takes the following form: consider a universality class composed of four different physical systems A-D. Each of A-D is described in step 1 by an effective Hamiltonian; effective Hamiltonians are ascribed to systems on the basis of various theoretical and empirical data. The RG explanation of universality, by virtue of steps 2-4, tells us that all the details which distinguish A-D, i.e. their irrelevant operators, are, in fact, irrelevant to the critical phenomena. Thus we have an explanation for how otherwise different systems exhibit the same phenomena at the critical point. This explanation relies, of course, on the RG transformations which allow for the categorisation of certain operators as irrelevant.

Importantly, this explanation takes the form of a general explanation of multiply realised phenomena: such phenomena are explained if commonalities are identified among the realisers and these are shown to be sufficient for the multiply realised phenomena to occur. Note that such explanations may be higher level and nothing written so far establishes their reducibility.

3 Anti-reductionist Arguments

Batterman (2000, 2017) and Morrison (2012, 2014) offer two arguments in defence of the view that the explanation just outlined is irreducible. The more general argument is that universality, *qua* instance of multiple realisability, is irreducible because multiple realisability requires abstracted explanations of a particular form.

However, one goal of this paper is to demonstrate that just such abstracted explanations may be reducible. Insofar as my reduction of the RG explanation goes through, we are thus faced with a dilemma: either some instances of MR are, in principle, reducible, or universality is not a case of MR. While I would opt for the former horn, nothing in the rest of the paper hangs on that choice.

The second anti-reductionist argument is much more specific to the case at hand and involves various demonstrations that the RG explanation requires infinities which are inexplicable from the bottom up. As noted by Palacios (2017), two different limits are invoked in the case of continuous phase transitions – the thermodynamic limit and the limit of scale invariance. There is an extensive literature on the thermodynamic limit as it appears in first order phase transitions; as I see no salient differences between appeal to this limit in the two contexts, I do not discuss this further here – see e.g. Butterfield and Bouatta (2012) for a reductionist account of that limit.[6]

The second limit is discussed by Butterfield and Bouatta (2012), Callender and Menon (2013), Palacios (2017), and Saatsi and Reutlinger (2018), among others, and these papers undermine claims that continuous phase transitions are irreducible. However, they pay insufficient attention to the specific role played by the RG (and by the limit of scale invariance) in establishing the irrelevance of certain details, and it is this role which is crucial to the anti-reductionist arguments.[7]

For Batterman, the RG is required because it allows us to answer the following question:

[6]The reductionist claims made here are conditional on a successful resolution of such issues.

[7]For example, Saatsi and Reutlinger (2018, p. 473) do not consider a counterfactual of the form 'if a physical system S did not exhibit effective scale invariance at criticality, then S would not exhibit the critical phenomena of any universality class' in their list of counterfactuals which the RG account is supposed to underwrite.

> **MR**: How can systems that are heterogeneous at some (typically) micro-scale exhibit the same pattern of behavior at the macro-scale? ...
>
> if one thinks (**MR**) is a legitimate scientific question, one needs to consider different explanatory strategies. The renormalization group and the theory of homogenization are just such strategies. They are inherently multi-scale. They are not bottom-up derivational explanations.
>
> [Batterman (2017, pp. 4, 14-15)]

As further elaborated below, the RG seems to Batterman to preclude "bottom-up derivational explanation" because it requires the following infinitary assumption:

> This [fixed point] is a point in the parameter space which, under τ [the RG transformation], is its own trajectory. That is, it represents a state of a system which is invariant under the renormalization group transformation. Of necessity, such a fixed point has an *infinite correlation length* and so lies on the critical surface S_∞. The singularity/divergence of the correlation length ξ is *necessary*.
>
> [Batterman (2011, p. 1045), original emphasis]

I accept that the RG formalism makes use of infinite limits. The salient question, to borrow Norton's (2012) distinction, is whether such infinities are approximations which allow one to use the more tractable infinitary mathematics to approximate features of the finite systems, or, alternatively, idealisations which describe a distinct infinite system. Claiming that the infinities are idealisations would preclude reduction because the macroscopic system with infinite properties has features which may not be reductively explained.

As Batterman demonstrates, the RG argument rests on the assumption of the infinite correlation length which generates absolute scale invariance. In §4 I claim that the physical systems under consideration are not absolutely scale invariant: in fact, one may abstract from the details of the underlying system insofar as such systems are effectively scale invariant; thus the infinitary assumption is best viewed as an approximation.

While Morrison (2014, p. 1155) likewise focusses on explanations of MR phenomena, she claims that RG explanations are irreducible for a different, but related, reason: the "RG functions not only as a calculational tool but as the source of physical information as well". Morrison (2012) makes a similar argument in relation to symmetry breaking in the physics of superconductors. She argues that, in both cases, top-down constraints play an essential role in the physical descriptions which thus rules out reduction. In the present context, Morrison's views may be understood as taking the RG invocation of scale symmetry to be a necessary physical assumption which cannot be understood from the bottom up. Below I argue that the effective scale invariance on which the RG rests is, in fact, reductively explicable. As such, no top-down organising principles are required and Morrison's claims are deflated.

4 Reducing the RG Explanation

Arguments for the reducibility of the explanation of universality have primarily been targeted at Batterman's claims that infinities are essential to the models used to describe continuous phase transitions. I do not have space to consider these arguments in any detail. Suffice it to say that, in my view, none succeeds in reducing the principal feature of the renormalisation group – the assumption of scale invariance. Thus I focus on that aspect of the RG, and claim that it, too, is reducible.

Furthermore, with the notable exception of Saatsi and Reutlinger (2018), not much attention has been paid to the explanation of universality *per se*. This, of course, makes a difference for MR-based objections to reduction, which raise doubts that a reductionist account could explain why the same phenomenon is exhibited in multiple different systems.

As far as the physics is currently developed, the RG plays an ineliminable role in the explanation of universality: it is the only mathematical framework available to predict the precise extent of observed universality of critical phenomena. If its application were truly mysterious, if we had no idea why it worked, then, infinity or no infinity, this would provide exactly the right kind of failure of explanation on which the anti-reductionist could hang their arguments.

I argue in the following that the applicability of the RG to systems un-

dergoing continuous phase transitions is not mysterious. The RG exploits effective scale invariance to set up equations which tell us how certain properties vary with respect to the variation of other properties. It is a piece of mathematics whose applicability is deeply physical – where the assumptions invoked in applying the RG do not hold, the RG's predictions go wrong.

In order fully to reduce the RG explanation, one also must consider the common features shared by each member of the same universality class, and argue that these, too, are reducible to aspects of the microphysical description. Such arguments have been given by the reductionists mentioned above. The innovation of this paper lies in reducing the RG framework, and the assumptions on which it relies; thus, given space constraints, I do not consider the reduction of the symmetry, dimensionality and representation by common Hamiltonians.

4.1 Reducing the Renormalisation Group

The RG argument rests on the assumption of scale invariance, and this is crucial to the demonstration that a class of operators are irrelevant at criticality. I claim that we can provide a bottom-up explanation of this scale invariance and that, as such, the RG arguments provide a mathematical apparatus for relating scale invariance to the irrelevance of certain details. One can see, heuristically, how scale invariance relates to universality: if the system at criticality is effectively scale invariant then many of that systems' features – those which are scale dependent – will turn out to be irrelevant at criticality, and all that will remain are those shared features such as the symmetry and dimensionality.

To argue that the RG explanation is reducible, I first give a more general characterisation of an RG flow. The calculation of each system's dynamics involves integration over a range of scales and energies. The highest energy (smallest scale) cutoff (denoted Λ) corresponds to the impossibility of fluctuations on a scale smaller than the distance between the particles in the physical system. The RG transformation involves decreasing the cutoff thereby increasing the minimum scale of fluctuations considered. Iterating this transformation generates a flow through parameter space designed to maintain the Hamiltonian form and qualitative properties of the system in question.

The RG transformation $\mathcal{R}$ transforms a set of (coupling) parameters $\{K\}$ to another set $\{K'\}$ such that $\mathcal{R}\{K\} = \{K'\}$. $\{K^*\}$ is the set of parameters which corresponds to a fixed point, defined such that $\mathcal{R}\{K^*\} = \{K^*\}$. This fixed point corresponds to the critical point defined physically. At the fixed point, the RG transformation (which changes the scale of fluctuations) makes no difference. Thus the fixed point encodes the property of scale invariance.

Given the Hamiltonian of one of our models, one can define an RG transformation which generates a flow that allows one to: (i) classify certain of the coupling parameters of the system in question as (ir)relevant to its behaviour near the fixed point, (ii) extract the critical exponents from the scaling behaviour near the fixed point.

The RG may be understood as a mathematical framework for exploring how certain properties vary with changing energy, length-scale, or, by proxy, temperature, on approach to the scale invariant critical point. Philosophical discussions of the RG are occasionally prone to mysterianism, but the RG should be considered to be no different from, for example, the calculus. As Wilson (1975, p. 674) notes: "the renormalization group ... is the tool that one uses to study the statistical continuum limit [the point of scale invariance] in the same way that the derivative is the basic procedure for studying the ordinary continuum limit".

The Hamiltonian which represents the system at the critical point, from which the critical exponents are extracted, is scale invariant at the fixed point – all the scale dependent contributions have gone to zero. Such Hamiltonians are known as 'renormalisable'. As such, the explanation provided below for the effective scale invariance of physical systems at criticality underlies the fact that such systems are well-described by renormalisable Hamiltonians at fixed points.

My argument has two steps: I demonstrate that scale invariance is implicit in the power law behaviour which is intrinsic to universality; then I provide a bottom-up explanation of the effective scale invariance for liquid-gas systems, a story somewhat motivated by the observation of critical opalescence. Thus, I show how scale invariance features in the mathematics – the Hamiltonian's renormalisability and the power laws, and how it features in the observed physics – the critical opalescence is a direct consequence of the bottom-up story.

The universality of critical phenomena lies in the sharing of power laws,

and hence critical exponents, between members of the same universality class. In what sense are such power laws scale-free? As Binney et al. (1992, p. 20) explain, a phenomenon obeying a power law is independent of scale because one could multiply its characteristic scale length by some factor and the ratio of values will remain constant. For example, consider the power law $f_1 = (r/r_0)^\eta$, and its measurement in the range $(0.5r_0, 2r_0)$. The ratio of largest to smallest value will be identical for measurements centred on $r_0, 10r_0, 100r_0$ – it will always be $4^{|\eta|}$, thus one may superimpose all the power laws by a simple change of scale. By contrast, for $f_2 = \exp(r/r_0)$ the ratio of values will change on scale changes.

Such systems are therefore described as scale-free; the RG is used to predict that at the point of scale invariance the heterogeneous features will be irrelevant. So, in order to work out when this framework is applicable, and why it works, we ought to look at each individual system, (for our purposes let's reserve inquiry to liquid-gas and ferromagnetic-paramagnetic systems) and identify the underlying processes which lead to effective scale invariance at the critical point. The following two caveats apply to this proposal for reduction:

First, it might be objected that universality may only be explained if the same processes are identified across all the systems exhibiting the universal behaviour; if that were so, the strategy employed here would be inadequate. However, universality may be explained by demonstrating that two conditions are fulfilled: that all the systems share common features, and that their heterogeneous details are irrelevant. While it's essential that the common features are shared by all the systems, the mechanism by which the heterogeneities are irrelevant may differ, so long as all the heterogeneities in fact end up as irrelevant.

Second, although the power laws and renormalisable Hamiltonians at the fixed point are absolutely scale invariant, the physical systems will, at best, be effectively scale invariant – that is, scale invariant within a certain range of length-scales. That should be acceptable because we know that scale invariance is never exactly true of a system: any real system will be finite and thus violate the assumption at some scale. Moreover, this will not generate empirical problems because the power laws are observed for systems approaching criticality – they are predictions about $T \to T_c$, not $T = T_c$. Thus one should only assume that critical exponents asymptotically approach those predicted at the fixed point. While infinite assumptions are required in order to impose the full scale invariance for RG analy-

sis, I claim that we can explain effective scale invariance for finite systems, and that absolute scale invariance is an approximation invoked to make the mathematics tractable.

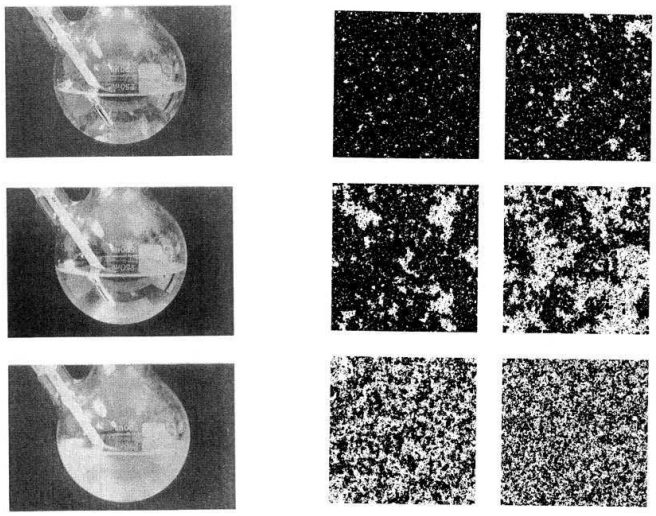
Scale invariance, as it manifests in systems at criticality, is known as 'self-similarity': as scales change the system resembles itself. How do we account for such self-similarity? The critical point, at which a continuous phase transition occurs, corresponds (for liquid-gas systems) to the highest temperature and pressure at which liquid and gas phases can be distinguished.

As is well known, there is a plateau in pressure-volume diagrams, which corresponds to the latent heat (or enthalpy) of vapourisation. This, roughly, is the extra energy needed to break the intermolecular bonds which distinguish liquids from gases and vapours. At the critical point this plateau, and the latent heat of vapourisation vanishes. Now it's difficult precisely to work out the binding energies of the intermolecular bonds. The values for this will be material dependent, and surface tension dependent, and will change at different pressures. But the heuristic argument tells us that the reason the plateau vanishes is because the system has enough temperature, and thus the molecules have sufficient energy to equal the binding energy. The point at which binding energy is exactly matched by kinetic energy will be the critical point.

The isothermal compressibility (κ) is defined as $\kappa_T = \frac{-1}{V}\left(\frac{\partial V}{\partial p}\right)_T$. This corresponds to how much the volume will change (∂V) with a given pressure change (∂p) at fixed temperature (T). As supercritical fluids have far higher compressibility than liquids, and both are present at the critical point, the compressibility diverges. Given, in addition, that the latent heat is zero at criticality, there's nothing to prevent a given bubble expanding arbitrarily. Thus we ought to expect the system to have bubbles of all sizes: this is what is meant by the claim that the system is dominated by fluctuations and has no characteristic scale.[8]

Negligible energy cost for transitions and infinite compressibility leads to self-similarity, and, in certain fluids, the bubbles at all scales lead to a high refraction of visible light. Thus otherwise transparent fluid may become opaque and milky-white. This is known as 'critical opalescence' – see figure 1(a) – and is a visible correlate of a system at criticality.

[8]Note that, for first order phase transitions, the compressibility also diverges; this doesn't lead to scale invariance because latent heat is finite.



(a) (b)

Figure 1: From Binney et al. (1992, pp. 10,19). (a) Critical opalescence is visible when arbitrarily large bubbles form in liquid at criticality. (b) Increasing loss of characteristic scale as $T \to T_c$ in simulations of the Ising model.

Such self-similarity is conceptually crucial to the applicability of the renormalisation group: in order to extract critical exponents from RG equations one identifies a renormalisable Hamiltonian which is scale invariant at the fixed point. Without fluctuations across all scales, systems would fail to be well modelled by such Hamiltonians. The physical argument for diverging fluctuation size justifies the use of a scale invariant mathematical model to represent such systems. Thus, for critical phenomena, the applicability of the RG depends on scale invariance, where this assumption is explicable from the bottom up.

Demonstrating these claims quantitatively is difficult, but the heuristic argument is convincing. Kathmann (2006) reviews theories of the nucleation of gas bubbles in water which generate accurate predictions concerning the rate of bubble growth and the threshold for stability over a range of temperatures; although these models do not reach the critical point, progress is being made.[9]

[9]Constructing exact models is especially difficult because of the fluctuations at a wide range of length scales – precisely the reason that the RG is employed.

Of course, further work could be done to develop these arguments and make them more precise. But there seems to be, in the above, a sound qualitative argument and no in-principle barriers to full derivation. This 'in-principle' ought not to be problematic: we know the relevant physical principles, even if quantitative models are still unavailable.

Moreover, as discussed below, and depicted in figure 1(b), the Ising model allows us quantitatively to predict analogues of the results for liquid-gas systems. While well short of a full explanation, the following discussion illustrates how self-similarity may be reduced for magnetic systems. By treating the Ising model as a stand-in for such systems, a similar kind of reasoning to that given above will go through.

Below the critical point, energy fluctuations will lead to random isolated spin flips. Such flips will be energetically costly and tend to be reversed. The higher the energy, the more likely these are to occur, and if sufficiently many occur then a patch will form, and other spins will have some tendency to align themselves with this patch. However, below the critical point, such patches beyond a certain size will be too costly and spins will overall remain aligned (there is some small probability of net magnetisation flipping, but this is increasingly unlikely further below the critical point).

At the critical point, the energy of the atoms in the lattice is greater than the energetic cost of violating spin alignment, and patches can become arbitrarily large. This results from the latent heat's vanishing and the divergence of the magnetic susceptibility (χ) on approach to the critical point. $\chi_T = \left(\frac{\partial m}{\partial B}\right)_T$ where m is the magnetisation and B represents an external magnetic field. Universality is manifested by the fact that the susceptibility and the compressibility both diverge according to identical power laws with the same critical exponent γ: χ_T, $\kappa_T \sim (T - T_c)^{-\gamma}$. Thus, we have self-similarity and effective scale invariance with bubbles or patches arbitrarily large up to the size of the system.

My aim is to establish the reducibility of the RG relevance and irrelevance arguments. I have demonstrated that the RG is a mathematical procedure that extracts information based on the empirically and theoretically justified assumption of effective scale invariance; this has been shown to be a property shared by different systems at criticality. The key ingredients for effective scale invariance are features of the interactions of neighbouring sub-systems, and the particulate constitution of the materials. While that suggests that these materials are not so different after all, it's worth empha-

sising that the systems which exhibit universal behaviour are nonetheless dissimilar away from the critical point – it's clear that magnets and liquids have many distinct chemical and physical properties.

The assumption of scale invariance plays a crucial role for the RG – it licences the discarding of scale dependent details; it is precisely this discarding of details which ensures that all systems are commonly described at the critical point. Moreover, discarding such details is what gives the higher-level explanation its stability and parsimony. It is thus incumbent on the reductionist to explain how the higher-level RG account is successful despite its leaving out such details. So, the reductionist should identify physical processes at the lower level which ensure the irrelevance of the discarded details.

As argued above, the physical processes in question are exactly those which lead to effective scale invariance. The fluctuations at all scales make it such that the scale-dependent properties which distinguish systems away from criticality are irrelevant at criticality, when the system is effectively scale invariant. We have identified, at the molecular level, the physical mechanisms which prevent variations in the discarded details from leading to changes in the higher-level description of the system. As such, we are assured that the explanatory value of the higher-level explanation is a consequence of features of the lower-level system.

One upshot of this reductionist account is that we may specify the conditions under which the higher-level description remains a good one. The discarded details are irrelevant while the large scale fluctuations – the bubbles or patches – dominate the physics. As we move to systems which are less scale invariant, as the bubbles die down, the critical point becomes a less accurate description and each system in the class will start to exhibit distinct behaviour. This is reflected in the fact that the macroscale RG description only derives the shared behaviour at the fixed point of scale invariance and predicts distinct behaviour away from the fixed point.

I end this section with the following intuitive physical gloss on the RG explanation: "[b]ecause the fluctuations extend over regions containing very many particles, the details of the particle interactions are irrelevant, and a great deal of similarity is found in the critical behavior of diverse systems" (A. L. Sengers, Hocken, and J. V. Sengers (1977, p.42)). Since we can explain the wide-ranging fluctuations from the bottom-up, the RG explanation of universality is reducible.

5 Conclusion

The field-theoretic RG framework, together with the common features of physical systems in the same universality class, explains how those systems all display the same critical phenomena when undergoing continuous phase transitions. That explanation is a higher-level explanation.

That higher-level RG explanation is nonetheless reducible. That is, we may explain in terms of the microstructure of each system how it is that each aspect of the higher-level explanation is explanatory. We may, in particular, show why the RG categorisation of operators as relevant and irrelevant works. That division depends on the assumption of scale invariance, and the assumption of scale invariance is justifiable when systems are effectively scale invariant at criticality.

The anti-reductionist claim that universality is MR, and MR is essentially irreducible has been undermined by demonstrating that we may arrive at a bottom-up understanding of the common features and of what makes such features sufficient for the common behaviour.

The further argument that the use of the infinite limit imposes an irreducible divide between the higher-level and lower-level models has similarly been countered: while we move to the infinite limit in order to make the mathematics simpler, the effective scale invariance can be shown to follow from details of the particle interactions at criticality – that's what identifies the critical point and allows us to make the corresponding abstractions from scale dependent details. Provided with this bottom-up explanation, there is no further reason to claim that the infinite limit is an idealisation rather than an approximation: for we have explained from the bottom up how the system is approximately self-similar.

One upshot of this discussion is that the RG is not to be regarded as mysterious, or, somehow, as the source of physical information. It is applicable only insofar as the systems to which it is applied have the relevant properties, and their having such properties may be reductively explained.

References

Batterman, Robert W. (2000). "Multiple Realizability and Universality". In: *The British Journal for the Philosophy of Science* 51.1, pp. 115–145.

Batterman, Robert W. (2011). "Emergence, singularities, and symmetry breaking". In: *Foundations of Physics* 41, pp. 1031–1050. DOI: 10.1007/s10701-010-9493-4.

— (2016). "Philosophical Implications of Kadanoff's work on the Renormalization Group". In: *Journal of Statistical Physics (Forthcoming)*.

— (2017). "Autonomy of Theories: An Explanatory Problem". In: *Noûs*. DOI: 10.1111/nous.12191.

Binney, James J. et al. (1992). *The Theory of Critical Phenomena: an Introduction to the Renormalization Group*. Clarendon Press, Oxford.

Butterfield, Jeremy and Nazim Bouatta (2012). "Emergence and Reduction Combined in Phase Transitions". In: *AIP Conference Proceedings* 1446, pp. 383–403. DOI: 10.1063/1.4728007.

Callender, Craig and Tarun Menon (2013). "Turn and Face the Strange ... Ch-ch-changes Philosophical Questions Raised by Phase Transitions". In: *The Oxford Handbook of Philosophy of Physics*. Ed. by Robert W. Batterman. Oxford University Press, pp. 189–223.

Fisher, Michael E. (1998). "Renormalization group theory: Its basis and formulation in statistical physics". In: *Reviews of Modern Physics* 70.2, p. 653.

Franklin, Alexander (2018). "On the Renormalization Group Explanation of Universality". In: *Philosophy of Science* 85.2. DOI: 10.1086/696812.

Kathmann, Shawn M. (2006). "Understanding the chemical physics of nucleation". In: *Theoretical Chemistry Accounts* 116.1, pp. 169–182. DOI: 10.1007/s00214-005-0018-8.

Mainwood, Paul (2006). "Is More Different? Emergent Properties in Physics". PhD thesis. University of Oxford.

Morrison, Margaret (2012). "Emergent Physics and Micro-Ontology". In: *Philosophy of Science* 79.1, pp. 141–166. DOI: 10.1086/663240.

— (2014). "Complex Systems and Renormalization Group Explanations". In: *Philosophy of Science* 81.5, pp. 1144–1156. DOI: 10.1086/677904.

Norton, John D. (2012). "Approximation and Idealization: Why the Difference Matters". In: *Philosophy of Science* 79.2, pp. 207–232.

Palacios, Patricia (2017). *Phase Transitions: A Challenge for Reductionism?* URL: philsci-archive.pitt.edu/13522/.

Saatsi, Juha and Alexander Reutlinger (2018). "Taking Reductionism to the Limit: How to Rebut the Antireductionist Argument from Infinite Limits". In: *Philosophy of Science* 85.3, pp. 455–482. DOI: 10.1086/697735.

Sengers, Anneke Levelt, Robert Hocken, and Jan V. Sengers (1977). "Critical-point universality and fluids". In: *Physics Today* 30.12, pp. 42–51.

Sober, Elliott (1999). "The multiple realizability argument against reductionism". In: *Philosophy of Science*, pp. 542–564.

Wilson, Kenneth G. (1975). "The renormalization group: Critical phenomena and the Kondo problem". In: *Reviews of Modern Physics* 47.4, pp. 773–840.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. Oxford University Press.

Title: There Are No Ahistorical Theories of Function

Author: Justin Garson

Abstract: Theories of function are conventionally divided up into historical and ahistorical ones. Proponents of ahistorical theories often cite the *ahistoricity* of their accounts as a major virtue. Here, I argue that none of the mainstream "ahistorical" accounts are actually ahistorical. All of them embed, implicitly or explicitly, an appeal to history. In Boorse's goal-contribution account, history is latent in the idea of statistical-typicality. In the propensity theory, history is implicit in the idea of a species' natural habitat. In the causal role theory, history is required for making sense of dysfunction. I elaborate some consequences for the functions debate.

Keywords: Philosophy of biology; biological function; selected effects; causal role; fitness contribution

Address: Department of Philosophy, Hunter College of the City University of New York, 695 Park Ave., New York, NY 10065

Email: jgarson@hunter.cuny.edu

1. Introduction

Theories of function are conventionally divided up into two main categories, historical and ahistorical (or backwards-looking and forwards-looking). The selected effects theory (Neander 1983, 1991; Millikan 1984) is an example of a *historical* theory, but there are other historical theories, including some versions of the organizational theory (McLaughlin 2001), and the weak etiological theory (Buller 1998). *Ahistorical* theories include Boorse's goal-contribution account (1976; 1977; 2002), the propensity theory (Bigelow and Pargetter 1987), and the causal role theory (Cummins 1975; Hardcastle 2002; Craver 2001; 2013). In the 1970s and 1980s, it was common to see these two sorts of theories as competing with each other, though more recently, philosophers of biology have generally adopted a pluralistic stance, and see them as capturing different aspects of real biological usage (OMITTED). Still, the validity of the basic distinction has never been seriously challenged.

Many proponents of ahistorical theories have argued that we should accept their theories precisely *on account of* their being ahistorical. In other words, their alleged ahistoriticity is often held up as a significant virtue of their theories, and a strong reason to prefer them to historical theories (or at least a strong reason to think they capture a significant strand of ordinary biological usage). There are two arguments along these lines. The first argument appeals to bald intuition, and says that it's just obvious that functions don't always need history. One fanciful variant of this argument appeals to science fiction cases, like swamp creatures, instant lions, and randomly-generated worlds (e.g, Boorse 1976, 74; Bigelow and Pargetter 1987, 188). But one doesn't have to go as far as science fiction to find plausible cases of ahistorical functions in biology. Many philosophers have a strong intuition that, the very first time a new biological trait emerges and begins to benefit the organism, it has a *function* even if it was never selected for (e.g., Boorse 2002, 66; Bigelow and Pargetter 1987, 195; Walsh and Ariew 1996, 498). The second argument, which is closely related, appeals to ordinary biological usage, not intuition. It says that historical theories run against the way biologists ordinarily think and talk about functions. At least sometimes, when biologists attribute functions to traits, they do not *cite* or *refer to* or *think about* history or evolution (e.g., Godfrey-Smith 1993, 200; Amundson and Lauder 1994, 451; Walsh 1996, 558; Boorse 2002, 73). Hence, ahistorical theories capture important strands of real biology.

In light of the above, my thesis might come as a bit of a shock. I claim that *there are no ahistorical theories of function* – or, to put it more precisely, the mainstream versions of the allegedly ahistorical theories on the market are not actually ahistorical. If we poke and prod at those theories a bit, a historical element falls out, like contraband stashed away in a suitcase. In Boorse's version of the goal-contribution account, history is explicitly embedded in his notion of a *statistically-typical* contribution to fitness. In the propensity account, history is embedded, a little less explicitly, in the idea of a species' *natural habitat*. Finally, I claim that the only way the causal-role theorist can hope to make sense of dysfunction is to appeal to history.

If this thesis is correct – that there are no ahistorical theories of function – three consequences immediately follow. First, we need to jettison this whole way of dividing up theories of function. The distinction between etiological and non-etiological theories serves us much better, as I'll describe in the conclusion. The distinction between etiological and non-etiological theories doesn't map onto the distinction between historical and ahistorical theories; rather, *these are two ways of being historical*. Second, given that there are no ahistorical views, a good portion of the arguments that have been put forward to date for these theories (those I mentioned above) are unsound. A third consequence is that one popular way of thinking about function pluralism must fail. This sort of pluralist wishes to sort all biological usage under two main umbrella theories, the selected effects theory and the causal role theory. An argument for this sort of pluralism is that it mirrors the two main uses of "function" in biology, the historical sense and the ahistorical sense. If I'm right, this incarnation of the pluralist project can't possibly work.

Before I move on, there is one big qualification I must get out of the way. One could, just for fun, *invent* a purely ahistorical theory of function. One could assert, for example, that *all* of a trait's effects are its functions. This theory (pan-functionalism?) would be ahistorical, to be sure, since even if the world were created two seconds ago in pretty much its present form, things would still have effects, and so they'd still have functions. In fact, sometimes scientists actually *do* use the word "function" synonymously with "effect." They say things like, "climate change is a *function* of deforestation," or "poor academic performance is a *function* of malnutrition." Clearly, there are some ahistorical uses of "function." But this isn't the ordinary biological use, which the theories I cite above are trying to capture.

So, I need to amend my thesis slightly. Instead of saying that there are no ahistorical theories of function, I want to say that any theory of function that satisfies two very minimal, very traditional, and largely uncontroversial, adequacy conditions, is *also* a historical theory. First, the theory should capture some distinction between functions and accidents (the function of the nose is to help us breathe but not hold up glasses). Second, the theory should capture the possibility of malfunctioning or dysfunction. If my heart seizes up due to cardiac arrest, it's failing to perform its function or it's dysfunctional. All of the theorists I engage with in this paper purport to satisfy these two adequacy criteria, or something like them, so I'm not begging any questions by insisting on these conditions.

Here's the plan for the rest of the paper. There are five sections. After the introduction, I'll turn to Boorse's version of the goal-contribution theory, and show how it explicitly contains a historical element (Section 2). Then I'll turn to the propensity theory and show how it contains a reference to history, buried inside the idea of a trait's *natural habitat* (Section 3). I will then show how the causal-role theory, if it is to make any sense of dysfunction, must include a reference to history (Section 4). In the conclusion (Section 5), I'll reiterate the big consequences for thinking about functions and suggest a better way of dividing up theories of function.

2. Boorse's Goal-Contribution Account

Seattle, WA; 1-4 November 2018 -269-

Boorse's view (1976; 1977; 2002), at the most general level, is a goal-contribution account. It holds that a trait's function is just its contribution to a goal. The plausibility of this view stems from its ability to reconcile artifact and biological functions in a single theory: the function of an artifact depends on its contribution to the goal of its user; the function of a biological trait depends on its contribution to the goal of the organism or the lineage. Here, I'll focus on the subclass of functions he calls *physiological* functions.

For Boorse, the *physiological* function of a trait is its species-typical contribution to the survival and reproductive prospects of an organism (1977, 555; 2002, 72). (To be more precise, Boorse carves up species into subgroups based on age and sex; the function of a trait is its typical contribution to fitness within the members of that subgroup.) Though he doesn't define a corresponding notion of *dysfunction*, he defines a closely related notion of *disease*: a disease is simply a state that "reduces one or more functional abilities below typical efficacy."

One of Boorse's arguments for the superiority of his theory over Wright's (1973) etiological approach, and the selected effects theory of Millikan (1984) and Neander (1983), is that his approach *makes no reference to history*. He advances two arguments for the value of this ahistorical approach; one appeals to ordinary biological usage, and the other appeals to intuition. First, he says, the goal-contribution account fits ordinary biological usage: "in talking of physiological functions, they [that is, pre-Darwinian biologists] did not mean to be making historical claims at all. They were simply describing the organization of a species as they found it" (1976, 74). The same is true of current physiologists, who have "*no thought* of explaining [a trait's] history" when they assign functions to them (Boorse 2002, 73, emphasis mine). All historical theories of function simply miss how physiologists have always used the word "function." His second argument appeals to intuition. He says that intuition revolts against putting history into functions, as attested to by his instant lions case. If the lion species sprang into existence by "unparalleled saltation," one would *not* say that the parts of lions don't have functions (ibid.; also see Boorse 2002, 75). Again, functions can't be historical.

Neander (1991, 182) raised a now-famous objection against Boorse; she pointed out that Boorse's view, as it stands, can't make sense of pandemic disease: "dysfunction can become widespread within a population…A statistical definition of biological norms implies that when a trait standardly fails to perform its function, its function ceases to be its function; so that if enough of us are stricken with disease (roughly, are dysfunctional) we cease to be diseased, which is nonsense." Pandemic diseases, moreover, don't just occupy the realm of science fiction, as in P. D. James' *The Children of Men*. UV radiation poisoning in anurans is a good example of pandemic dysfunction. Sadly, climate change might create many more pandemic dysfunctions very soon. A good theory of function should at least allow for the *conceptual* possibility that all, or most, tokens of a certain trait in a certain species are dysfunctional (or as Boorse prefers, "diseased").

Intriguingly, Boorse doesn't deny the possibility of pandemic disease. Instead, he says that in order to make sense of pandemic disease, one has to appreciate function's

historical depth. Specifically, he says that when we consider what is "statistically typical" for a trait, we cannot just look at what is typical right now. Rather, we have to consider what is typical within a long slice of time that extends far back into the past: "Obviously, some of the species' history must be included in what is species-typical. If the whole earth went dark for two days and most human beings could not see anything, it would be absurd to say that vision ceased to be a normal function of the human eye (2002, 99)." He tells us that this time-slice should be longer than "a lifetime or two," and might include "millennia."

This is an extraordinary admission, given that much of Boorse's core argument *for* his view was propped up on the claim that both biology and intuition need purely ahistorical functions, uncluttered by history. His admission implies that two of his key arguments for the view (cited above), are unsound. First, by his own admission, it's not the case that biologists don't refer to history; implicitly, when they talk about what's statistically-typical, they *are* talking about history. Second, regardless of whether or not intuition supports ahistorical functions, Boorse's theory doesn't. It's just not true, on Boorse's account, that if lions popped into being from an unparalleled saltation, their parts and processes would have functions. They wouldn't, since they don't have the right history (or to be more precise, they have no history at all). True, Boorse's history isn't the same *kind* of history that features in the selected effects theory, since it doesn't refer specifically to etiology, but it's still history, and so his arguments that appeal to the ahistoricity of his theory don't work.

3. The Propensity Theory

Bigelow and Pargetter (1987) also developed an influential "ahistorical" theory of function, the propensity theory. They reject the selected effects theory (and etiological accounts more generally) because the selected effects theory gets the *modality* of functions wrong. In other words, the statement, "functions are selected effects," if true, is contingently true; it might be true on the actual world, but there are possible worlds at which it's false. To illustrate the point, they ask us to consider a world that is pretty much the same as ours except that it randomly popped into being five minutes ago. On that world, they claim, there would still be functions, just no selected effects (188): "we have the intuition that the concept of biological function…[is] not thus contingent upon the acceptance of the theory of evolution by natural selection." This consideration prompts the need for an ahistorical theory.

For Bigelow and Pargetter, functions are propensities, or probabilistic dispositions. We might quibble over what exactly dispositions are, but any good definition will cite three parts: structure, environment, and behavior. Consider the solubility of salt. There is a *structure*, namely, the polar molecular structure composed of sodium and chloride; there is an *environment*, namely, water; there is a *behavior*, namely, dissolving. When we say that salt is disposed to dissolve in water, we're saying that, if you were to take this structure, and put it in this environment, it would perform this behavior.

Functions, too, are dispositions. Consider "the function of the heart is to circulate blood." For this statement to be true, there must be a structure (the heart, embedded the right way in the circulatory system), an environment (which they call the creature's *natural habitat*), and a behavior (conferring a fitness boost on the organism). If one were to put the structure in its natural habitat, it would increase the fitness of the organism (relative, I suppose, to creatures without hearts). The crucial distinction between their view and Boorse's is that in their view, a trait's function doesn't depend on actual frequencies of performance. A trait needn't have an actual track record of boosting fitness to have a function; a mere propensity will do.

This raises the thorny question of what a creature's *natural habitat* is. For they're clear that a creature's natural habitat isn't just any environment the creature happens to find itself in. Unfortunately, they refuse to define this crucial notion; instead, they brush it off as vague, but unproblematically so: "there may be room for disagreement about what counts as a creature's 'natural habitat;' but this sort of variable parameter is a common feature of many useful scientific concepts" (192). But one could at least form the suspicion that if one analyzed this unproblematically vague notion, one would find some reference to history tucked away inside of it.

This suspicion is confirmed in the very next paragraph. There, they tell us that, if a creature's environment were to change very suddenly, then "natural habitat" will still refer to the *old* environment, and not the *new* one (ibid). There's a time lag built into the very idea of a natural habitat. So, for example, if climate change melts enough Arctic ice, then, at least for a time, the polar bear's natural habitat (and by extension, the natural habitat of the trait itself, namely, their thick, water-repellant fur) is the icy habitat of yore and not the contemporary, denuded one. They take that as given, and I agree.

But why would this be? What *makes it the case* that this is true, namely, that in cases of rapid habitat change, "natural habitat," at least for a time, refers to the old environment and not the new one? What makes it true, I suspect, is that the idea of a natural habitat is an intrinsically historical notion. It's something like the environment within which the organism recently survived and thrived. And if that's not what a natural habitat is, I would like to know what it is *such that*, if a creature's actual habitat shifts suddenly, the natural habitat is still the old one. Just because a concept is vague around the edges, that doesn't exempt one from the obligation to give some sort of analysis.

Hence, I conclude that, contrary to rumor, the propensity theory is not an ahistorical theory, or not demonstrably so. But if that's right, they lose one of the main virtues of the view, which is to get the modality of functions right. To be fair, there's still a sense in which their view *is* ahistorical. What they can do, that the selected effects theorist can't, is to attribute functions to novel traits – so long as that novel trait belongs to the members of a species that has been around long enough to have a natural habitat. Suppose a gene mutation confers a benefit on an organism, say, pesticide resistance on a flour beetle. I suppose they can say that, at the very moment at which it first confers that benefit, the gene mutation has a function, namely, to make the beetle withstand a certain pesticide. This result, they claim, is "intuitively comfortable" (195). But they can say that only

because flour beetles themselves have a history, and so we can talk meaningfully about their natural habitats. Moreover, I think they'll still have a very hard time dealing with dysfunction (Neander 1991, 183), as I hope to show in the next section. Finally, I think there are good theory-neutral reasons for saying that beneficial traits, on their very first appearance, don't have functions, but rather, whatever benefit they bring is an accident. But I won't argue for that here (see OMITTED).

4. The Causal Role Theory

What about the causal role theory of function? This appears to be a purely ahistorical view. The causal role theory says, roughly, that the function of a *component* of a system consists in its contribution, in tandem with the other components, to a system-level capacity of interest (Cummins 1975; Craver 2001; Hardcastle 2002). Craver (2001; 2013) helpfully elaborates this view by specifying that the part in question must be a component of a *mechanism*. All of the basic ingredients of this theory are ahistorical: capacities, components, organization, hierarchy, interests. Even if the world were created five minutes ago, in pretty much its present form, things would still have causal role functions.

The problem enters when we think about dysfunction. Cummins (1975, 758) insisted that functions are dispositions, or capacities: "…to attribute a function to something is, in part, to attribute a disposition to it." The function of a trait *token*, then, consists in its capacity to contribute to a system-level effect. But what if the token in question, through defect or disease, loses the capacity, and so can't contribute to the system-level effect? Then, by Cummins' analysis, it doesn't have the relevant function – so it can't dysfunction either.

Causal role theorists have, by and large, been silent about how to make sense of dysfunctions from this perspective. Almost everything they've had to say on that score, however, is consistent with the following theme: a trait *token* dysfunctions when it can't do what other trait tokens generally, or typically, do to contribute to the system-level effect of interest. Consider Godfrey-Smith (1993, 200): "Although it is not always appreciated, the distinction between function and *mal*function can be made within Cummins' framework…If a token of a component of a system is not able to do whatever it is that other tokens do, that plays a distinguished role in the explanation of the capacities of the broader system, then that token component is malfunctional." Craver (2001, 72), offers the same general line: "…the ascription of a function to a malformed or broken part is derivative upon a description of how that *type* of part (X) fits into a *type* of higher-level mechanism (S). The malformed and broken part can be identified as an X by the typical properties and activities of Xs...." This is, at root, to rely on a statistical norm for making sense of dysfunction.

This account of dysfunction, like Boorse's, stumbles when it encounters the problem of pandemic dysfunction (Neander 1991). For the modification suggested above implies that, if everyone's heart seized up at once, nobody's heart would have a function anymore, so nobody's heart would be dysfunctional. The best way to solve this problem,

and perhaps the only way, is the way Boorse took, namely, to say that the function of a trait is its typical contribution to some system effect, when what's typical is assessed over a chunk of time that stretches back into the past, for at least "a lifetime or two," and perhaps "millennia." But if causal role theorists take that line, they'd have a historical theory.

Craver (2001) and Hardcastle (2002) suggest, all too fleetingly, a different way of thinking about dysfunction, one that depends not on statistics, but on our values and goals, that is, the values and goals of people who make function attributions. Craver (2001, 72) suggests that traits dysfunction when they cannot do what people *want* them to do: "the mechanistic role of the broken part only appears against the fixed backdrop of shared assumptions about a type of mechanism within which parts of this type generally (or preferably) make important contributions." The parenthetical remark alludes to a substantially new doctrine, one that demands our full concentration. It suggests that dysfunction is a mirror of human preferences and goals, of our wishing and wanting. If my heart seizes up, it's dysfunctional, since it's not doing *what I want it to do.*

Hardcastle (2002) makes remarks along similar lines. She first says that the function of a trait - what it's "supposed to do," as she puts it - depends on the goals of the scientific discipline that makes the investigation: "The teleological goal for some trait…depends upon the discipline generating the inquiry" (153). The palmomental reflex causes a chin twitch when you stroke an infant's palm; it's just an accident of cortical wiring with no deep evolutionary rationale. Still, she says, it has the *function* of indicating the state of brain development in infants, because that's how biomedical researches use it. She then says that something malfunctions just when it cannot do what it's supposed to do (152). The palmomental reflex malfunctions when it can't indicate the state of brain development. Simply put, dysfunction happens when a trait can't do what we want.

But dysfunctions cannot be reduced to preferences in any straightforward way; this is a point that's been taken for decades (e.g., Boorse 1977, 544; Wakefield 1992, 372), for reasons that scarcely need to be rehearsed. I'd prefer not to need sleep and water; I'd prefer if nobody had to go through the pain of childbirth or teething, either. But none of those things are diseases or dysfunctions. For that matter, I'd prefer if my hands were equipped with retractable adamantium claws. The fact that my hands can't do what I want them to do doesn't make them dysfunctional. If one really wanted to run with this value-centered line about dysfunction, one would *at least* have to add that, in order for a trait to dysfunction, it's not enough that it doesn't do what I prefer, but I must also have a *reasonable expectation* that it *should* act in the way that I prefer. But what could possibly ground a *reasonable expectation* that my hand (say) work in a certain way? Only this: that hands usually *do* work in the preferred way. But then we're back to statistical norms, and long historical slices of time. This value analysis of dysfunction isn't a contender to a statistical analysis; instead, the former presupposes the latter.

I've walked through three allegedly ahistorical theories of function, and shown that none of them are purely ahistorical; they're tainted with history. The conclusion will say what we should do next.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -274-

5. Conclusion

There are no ahistorical theories of function, at least among those that are usually put forward as ahistorical. The first, Boorse's goal-contribution theory, explicitly refers to what is statistically typical for a trait, where what's typical is assessed over a long historical period of time. The second, the propensity theory, refers to the creature's natural habitat, which is implicitly historical. And the third, the causal role theory, can't hope to make sense of dysfunction (or so I argue) without appealing to a statistical norm, and thereby (following Boorse) to history. *No* theory of function will give functions to the parts of swamp creatures, instant lions, or anything on worlds that are similar to ours except for being randomly generated five minutes ago. The propensity theory, at least, can give functions to novel traits as soon as those traits begin benefiting their bearers, as long as the population in which the traits emerge has been around for long enough to have something like a natural habitat. But even that theory will probably encounter problems when it comes to making sense of dysfunction, though I haven't pushed that line in any detail here.

Three immediate consequences follow from this fact. The first is that we should stop dividing up theories of function in terms of historical and ahistorical. The second is that many of the main arguments for the allegedly ahistorical theories are unsound. Third, one popular form of pluralism, which says that there are two main theories of function, corresponding to historical and ahistorical uses of "function" in biology, is untenable.

But if we can't rely on the historical/ahistorical distinction as a way of dividing up functions, how should we talk about them? I think it's best to divide them up into etiological and non-etiological (as theorists are sometimes wont to do anyway). But there's a crucial clarification in order: to say a theory is etiological isn't *just* to say it's historical. It's to say that the theory deals specifically with causal history. The theory purports to capture the sense in which, when we attribute a function to a trait, we're trying to give a causal explanation for why the trait exists. Most other theories of function are non-etiological, in that they do not purport to explain, in a causal sense of "explain," why the trait exists. But they're still historical.

There's a twist to this story. I think there *are* ahistorical theories of function. Consider that climate change is a function of deforestation, poor academic performance is a function of malnutrition, and wildlife habitat is a function of soil. These notions are *ahistorical* through and through. "Function," in this context, means little more than "effect," and perhaps (as in the last of the three examples) "helpful effect." But this tepid sense of function isn't going to sustain a distinction between function and accident, nor will it give us any sense of dysfunction. This is the sort of "function" that Bock and von Wahlert (1965, 274) were getting at when they equated functions with "all physical and chemical properties arising from [the trait's] form." It's also the sort of "function" that Neander (2017) describes in her recent discussion of "minimal functions." But the proponents of the allegedly ahistorical theories want functions to do much more than that. They are trying to capture the ordinary biological sense (or *an* ordinary biological sense)

of "function," where functions differ from accidents and sometimes things dysfunction. Unfortunately, they can't have what they want.

References

Amundson, R., and G. V. Lauder. 1994. Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy* 9: 443-469.

Bigelow, J., and Pargetter, R. 1987. Functions. *Journal of Philosophy* 84: 181-196.

Bock, W. J., and von Wahlert, G. 1965. Adaptation and the form-function complex. *Evolution* 19: 269-299.

Boorse, C. 1976. Wright on functions. *Philosophical Review* 85: 70-86.

Boorse, C. 1977. Health as a theoretical concept. *Philosophy of Science* 44: 542- 573.

Boorse, C. 2002. A rebuttal on functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M. Perlman, 63-112. Oxford: Oxford University Press.

Buller, D. J. 1998. Etiological theories of function: A geographical survey. *Biology and Philosophy* 13: 505-527.

Craver, C. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Craver, C. 2013. Functions and mechanisms: A perspectivalist view. In *Function: Selection and Mechanisms*, ed. P. Huneman, 133-158. Dordrecht: Springer.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72: 741–765.

Godfrey-Smith, P. 1993. Functions: Consensus without unity. *Pacific Philosophical Quarterly* 74: 196-208.

Hardcastle, V.G. 2002. On the normativity of functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M Perlman, 144-156. Oxford: Oxford University Press.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Neander, K. 1983. *Abnormal Psychobiology*. Dissertation, La Trobe.

Neander, K. 1991. Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.

Neander, K. 2017. Functional analysis and the species design. *Synthese* 194: 1147-1168.

Seattle, WA; 1-4 November 2018 -277-

Wakefield, J. C. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47: 373–388.

Walsh, D.M. 1996. Fitness and function. *British Journal for the Philosophy of Science* 47: 553-574.

Walsh, D. M., and A. Ariew. 1996. A taxonomy of functions. *Canadian Journal of Philosophy* 26: 493-514.

Wright, L. 1973. Functions. *Philosophical Review* 82: 139-168.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -278-

What do molecular biologists mean when they say 'structure determines function'?

Gregor P. Greslehner*

University of Salzburg & ERC IDEM, ImmunoConcept, CNRS/University of Bordeaux

October 2018

Abstract

'Structure' and 'function' are both ambiguous terms. Discriminating different meanings of these terms sheds light on research and explanatory practice in molecular biology, as well as clarifying central theoretical concepts in the life sciences like the sequence–structure–function relationship and its corresponding scientific "dogmas".

The overall project is to answer three questions, primarily with respect to proteins: (1) What is structure? (2) What is function? (3) What is the relation between structure and function?

The results of addressing these questions lead to an answer to the title question, what the statement 'structure determines function' means.

*Email: gregor.greslehner@gmail.com

Keywords: philosophy of biology, molecular biology, protein structure, biological function, scientific practice

1 Introduction

'Structure' and 'function' are abundantly used terms in biological findings. Frequently, the conjunct phrase 'structure *and* function' or the directional phrase '*from* structure *to* function' is to be found, indicating that there is a special relation connecting these two concepts. The strongest form of this relation is found in the frequent statement that 'structure *determines* function'. One could easily list several hundreds of references containing such phrases. However, in order not to blow up the references section, I will refrain from doing so. Suffice it to say that biologists make highly prominent use of these concepts in describing their research—molecular biologists, in particular. In this paper, I attempt to clarify these concepts, address their relation, and discuss the role they play in molecular biology's explanatory practice. While these issues can be addressed for many different biological entities on different levels of organization, I restrict the discussion primarily to proteins.

What do biologists refer to when they use this phrase? Is there a particular scientific program or strategy behind the slogan 'structure determines function'? Despite the frequent use of this phrase and the concepts to which it refers, a rigorous analysis is missing. Thus, a philosophical clarification would be a valuable contribution to the conceptual foundations of biology. One such fundamental concept is the sequence–structure–function relationship. "The relationships between sequence, structure, biochemical function and biological role are extremely ill-defined and scant

high quality data are available to allow us to analyse them." (Sadowski and Jones, 2009, 360)

In this paper, I attempt to close this gap by developing an explication of both concepts of *structure* and *function* as they are used in biological practice and discussing which relation holds between them. The third component in this "trinity of molecular biology"—sequence—is the least in need of explication. The standard textbook view holds that sequence determines structure, and structure determines function. I will focus on the second relation.

Without reviewing the rich history of these concepts throughout biology at this point, it is worth noting that functionality and form or structure were thought to be intimately linked from early on. In the early days of biology at the macroscale, the structures had to be observed with the naked eye. Thus, the first examples about the form of bodies or their parts and their functions can be found in physiology and anatomy, for example Harvey's notion of the heart's function to pump blood. From the scale of physiology to the molecular scale, structure and function are closely related. What exactly links these two concepts? Is it a determination relation? And if so, which one is determining the other?

With the invention of microscopes and later the emergence of molecular biology, the structures and functions under consideration shifted from macroscopic entities to individual molecules. In fact, molecular biology put the three-dimensional shape of molecules center stage for explaining biological phenomena. This is the focus of this paper. In particular, the discussion will be confined to the structure and function of *proteins*—with special emphasis on the question whether the former determines the latter.

2 The ambiguity of 'structure'

In a first approximation, 'structure' and 'function' could be interpreted as the most general or neutral way of describing what molecular biologists are doing in their research and what their findings are about. These include mainly the three-dimensional shapes of molecules (or larger cellular structures) and the activities (functions) theses molecules perform in living cells, biochemical pathways, chemical reactions, or just individual steps in such mechanisms. The ultimate aim is to explain biological phenomena with molecular mechanisms, whose entities can be described in physical and chemical terms. The structure of molecules can be described in terms of physics and chemistry—function, however, is a concept that does not appear in physics or chemistry. Let's start by taking a closer look at the notion of structure.

'Structure' is an ambiguous term. Applied to proteins, there is the usual nomenclature of *primary structure* (i.e., a protein's amino acid sequence), *secondary structure* (i.e., common structural motifs like α-helices and β-sheets), *tertiary structure* (i.e., the three-dimensional shape of a single folded amino acid chain), and *quaternary structure* (i.e., the final assembly of a protein if it consists of more than one amino acid chain). Other structurally important components are post-translational modifications and prosthetic groups which are not part of its amino acid composition. All these notions of structure have in common that they are about the molecular composition and shape of a molecule. One meaning of 'structure' denotes the sequence of a polymer, the other meaning is about the three-dimensional shape of a molecule. As will be discussed below, another important ambiguity of 'structure' allows to denote the organization of an interaction network. That leaves us with three different meanings of 'structure':

(1) the sequence of a polymer, (2) the three-dimensional shape of a molecule, and (3) the network organization of several biological entities.

While meanings (2) and (3) are candidates for being functional entities, structure as sequence (1) rather relates the sequences of different polymers (DNA, RNA, and proteins) and also plays a central role in determining the three-dimensional shape of a molecule, structure (2). The primary structure of a protein is just the sequence of amino acids that are put together to form a polypeptide. This amino acid sequence is determined by the corresponding protein-coding gene, which is first transcribed into mRNA and then translated into protein by the ribosome. This scheme is known as the "central dogma of molecular biology":

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$$

The arrows might be interpreted as determination relations. The textbook view of protein structure and function proceeds as follows:

$$\text{nucleotide sequence} \rightarrow \text{amino acid sequence} \rightarrow \text{protein structure} \rightarrow \text{protein function}$$

Strong evidence supporting the claim that the three-dimensional shape of a protein is determined by the sequence of amino acids alone was provided by the experiments of Christian Anfinsen, showing that ribonuclease could, after treatment with denaturing conditions, regain its form and function (Anfinsen et al., 1961). Later, Merrifield showed that an *in vitro* synthesized sequence of amino acids can carry out the enzymatic activity of ribonuclease, thus gaining its functional form without the aid of any other cellular component (Gutte and Merrifield, 1971). From this and similar experiments, Anfinsen

Seattle, WA; 1-4 November 2018 -283-

built general rules of protein folding as a global energy minimum which depends solely on the sequence of amino acids (Anfinsen, 1973). This view is known as "Anfinsen's dogma".

In 1958, John Kendrew's lab determined the first actual three-dimensional form of a protein, myoglobin (Kendrew et al., 1958). The predominant technique to determine protein structures is still X-ray crystallography (Mitchell and Gronenborn, 2017). Other techniques include nuclear magnetic resonance, cryogenic electron microscopy, and atomic force microscopy. X-ray structures in particular have been supporting the view that there is a unique rigid shape—the protein's native, functional state—which would be necessary and sufficient for a protein to carry out its biological function.

To make a long story short, the relation between nucleotide sequence and amino acid sequence has been generally confirmed (although there are much more complicated mechanisms to it, e.g., splicing). However, the part concerning the protein shape and function proves to be much more problematic. That poses a challenge to what Michel Morange calls "the protein side of the central dogma" (Morange, 2006).

To get from amino acid sequence to three-dimensional structure is known as the *protein folding problem*. As the term 'problem' suggests, it poses a serious challenge and remains unsolved to this day. Even though knowledge-based techniques to predict protein structures from their sequence have become impressively sophisticated, successful, and reliable, there are good reasons to suspect that the protein problem might remain unsolved in principle—if the aim is to predict protein folding based on chemical and physical principles only.

Every two years the best prediction tools are tested in a contest, the Critical Assessment of protein Structure Prediction (CASP). Based on experimentally determined structures which are only published after the participants of the contest have

submitted their predictions, the predictions are then compared to the experimental structure. A similar contest for predicting the functions of proteins exists (Critical Assessment of Functional Annotation, CAFA), although it is much less developed. But what is function in the first place?

3 The ambiguity of 'function'

'Function' is also an ambiguous term (Millikan, 1989)—even more so than 'structure'. There is a rich history of debates surrounding different notions of function. The term 'function' has a long tradition in biology and its philosophy (Allen, 2009). Starting with Aristotle, activities in biology were interpreted to *have a purpose*, to be goal-directed (teleological). The standard example is that the heart's function is to pump blood. That the heart also produces noise is not considered to be functional. Classic accounts of function have been predominantly trying to capture the teleological aspect, for example (Wright, 1973). However, intentionality is a problematic notion in biology. In another important account, Robert Cummins (1975) stressed the importance of a component's contribution to the system in which it is contained, rather than why natural selection has favored a certain trait. Although it makes sense in evolutionary biology to have an account of function that captures the evolutionary developments, molecular biology and protein science operate with a different notion of function, i.e., mainly biochemical activity. There seem to be two entirely different questions: What is a structure doing? And how did this structure evolve to do what it does?

Arno Wouters distinguishes four notions of biological function (Wouters, 2003): (1) (mere) activity, (2) biological role, (3) biological advantage, and (4) selected effect.

The last two are issues of evolutionary biology, whereas the former two fall within the molecular biologist's domain. If function is to be determined by a molecule's three-dimensional shape or organization network, only (1) and (2) seem to be the proper reading of 'function' in this context.

Which entities have functions within living organisms? Depending on the level of organization at which one is operating, one could give a different answer: molecules, organelles, cells, tissues, organisms, individuals, populations, ecosystems. The most prevalent candidates in molecular biology are certainly DNA and proteins, although lipids and other biomolecules play important roles in life processes, too.

Traditionally, functions have been attributed to entire genes ("one gene—one enzyme hypothesis"). These views are related to the genetic determinsim view of having a gene for every trait, in which every gene has a function. However, the primary functional units inside a cell are arguably its proteins. Their biochemical activities and biological roles depend crucially on their three-dimensional shapes and network organization, respectively.

One has also to take into account more abstract functional entities, i.e., network modules. These are also called 'structures' but do not refer to the shape of molecules. Its functions ought to be considered as Wouter's second notion (biological role), rather than biochemical activity. "Current 'systems' thinking attributes primary functional significance to the collective properties of molecular networks rather than to the individual properties of component molecules" (Shapiro, 2011, 129). "[A] discrete biological function can only rarely be attributed to an individual molecule [...]. In contrast, most biological functions arise from interactions among many components." (Hartwell et al., 1999, C47). Thus, we can attribute functions as biochemical activities to

individual molecules, wheres systems functions (biological roles) are attributed to organizational structures:

> "Finding a sequence motif (e.g., a kinase domain) in a new protein sheds light on its biochemical function; similarly, finding a network motif in a new network may help explain what systems-level function the network performs, and how it performs it." (Alon, 2003, 1867)

4 Does structure determine function?

Having distinguished between three notions of 'structure' and two notions of 'function', what about the statement 'structure determines function'? Is—in any of its different readings—a certain structure necessary or sufficient for a certain function?

The common textbook view according to Anfinsen has a clear answer: "the central dogma of structural biology is that a folded protein structure is necessary for biological function" (Wright and Dyson, 1999, 322). On first glance, it might appear plausible to assume that a particular structure (understood as molecular shape) is a necessary condition for the proper function of a biological structure (i.e., its biochemical activity). Loss of function is often associated with a loss of the three-dimensional shape of individual proteins. On the other hand, to go for the "sufficient" direction, changes in structures often lead to a decrease in functionality, up to a complete loss. Many diseases for which there are known molecular causes give support to this view. Often it is alterations in the sequence of DNA that result in changed protein shapes that lead to a functionality defect of the organism, which is the definition of a "molecular disease". Alterations of a protein's three-dimensional shape, however, do not necessarily lead to

loss of function. In many cases, changes are "silent", i.e., they don't cause any alteration in phenotype. In rare events, changes might even turn out to be "improvements", which is the driving force of evolutionary development.

However, evidence has been found in the recent years that a significant portion of proteins are intrinsically unstructured in order to be functional, see for example (Forman-Kay and Mittag, 2013). Does the discovery of intrinsically unstructured proteins challenge the relation between structure and function? "[D]isorder aficionados are calling for a complete reassessment of the structure-function paradigm" (Chouard, 2011, 151). Some protein domains fold only upon binding to a suitable target. Others, however, seem to never have an ordered state at all—they remain unstructured even in their functional state.

That a high similarity in sequence does not guarantee a similarity in structure or function has been shown by the Paracelsus Challenge: "a one-time prize of $1000, to be awarded to the first individual or group that successfully transforms one globular protein's conformation into another by changing no more than half the sequence" (Rose and Creamer, 1994, 3). One recent answer to this challenge resulted in the synthesis of two proteins which have 88% sequence identity but a different structure and a different function (Alexander et al., 2007).

Contrary to the view described above, the generalization that a stable three-dimensional structure is necessary or sufficient for a particular function does not hold. It remains true, however, that there is an intimate correlation between structure and function. Prediction tools based on this view are a powerful tool. An attempt to systematically predict the structure and function of proteins based on their amino acid sequence can be found, for example, in (Roy et al., 2010).

To complicate the picture, codon usage is also important: Zhou et al. (2013) have shown that the FRQ protein, which is involved in the circadian clock, is using non-optimal codons, thus translation speed is not optimal. After experimentally optimizing codon usage, the resulting protein—which has the exact same amino acid sequence—folds differently and is no longer functional. This shows that amino acid sequence by itself is not sufficient to determine the three-dimensional structure, let alone its function. In addition to the correct sequence, the folding process has to take place in a certain way which is influenced by the usage of codons and thus the availability of tRNAs, which influences the speed at which the ribosome can proceed translation. Usage of non-optimal codons gives the nascent polypeptide chain some time for the segments that have already been translated to fold in a certain conformation. If translation is too fast, certain intermediate folds which are necessary to reach the final functional conformation can be lost.

Another idea to keep in mind is that evolution operates pragmatically: structures are not the target of selection, functions are. Structures are being re-used for novel functions—there are many biological examples.

If structure does not *determine* function, if a particular structure (in any of its three meanings) is neither necessary nor sufficient for a particular function (in any of its two meanings), may there be another way in which structure and function are related? Perhaps there is a less stringent relationship? I will argue for a supervenience relation (McLaughlin and Bennett, 2018). But before developing this account, we need to clarify which notions of 'structure' and 'function' to use to capture actual scientific practice in molecular biology.

In order to speak about biological functions, a reglemented vocabulary is needed.

The most successful of these is gene ontology (GO) (Ashburner et al., 2000). Fascinating correlation analysis between three-dimensional protein structures from the Protein Data Bank (PDB) and GO terms can be found, for example, in (Hvidsten et al., 2009) and (Pal and Eisenberg, 2005).

According to the textbook picture, there is a linear chain of determination, leading from nucleotide sequences in the DNA via transcription to the nucleotide sequence of RNA, which leads via translation to the amino acid sequence of proteins. The sequence of amino acids, in turn, determines the three-dimensional structure of the protein, whose function, again, is determined by its structure. Given transitivity of this determination relation, one would only need to know the genomic sequence in order to have a complete picture ("blue print") of the functional organism. That is the "holy grail of molecular biology". And like the quest for the holy grail, it is doomed to fail. A strict determination relation does not even hold between the individual pairs.

The reason why the simplified scheme above is still part of the current research "paradigm" lies, on the one hand, in its scientific success: genomics and proteomics have provided unimaginable insights. On the other hand, it fits the mechanistic, reductionistic narrative that has been fashionable in molecular biology. Today, systems biology claims to provide a "holistic" alternative (Green, 2017).

But even without such a strict determination relation between structure and function, both concepts are central to explaining molecular mechanisms in research practice.

In order to understand why molecular biologists explain mechanisms with reference to structure and function, we need to understand what these concepts denote. In a first approximation, molecular biologists analyze a phenomenon by identifying its components that are responsible for the phenomenon in question. These components are the

structures that perform certain biochemical activities, which collectively bring about the phenomenon (biological role). The way in which these entities and their activities are organized is a different meaning of 'structure' which is as important in a mechanistic explanation as individual molecular structures are.

> "Despite the lack of an overarching theory, a Newtonian or quantum mechanics of its very own, molecular biology has become a unifying discipline in virtue of the powers of its techniques, its ability to extrapolate from the molecular to higher levels, and its synthesis of problems of form and function at the molecular level. This synthesis of form and function is a central, ill-understood, and historically important feature of molecular biology." (Burian, 1996, 68)

The ambiguity of the terms 'structure' and 'function' might be useful, for it can be applied to a broad variety of biological research strategies and activities. But, on the other hand, using the term same for different things causes confusion, and the use of metaphorical language might be obscuring certain features and difficulties with this approach.

More recent and thriving approaches in the life sciences have moved beyond the idea that there is a determination relation between structure and function and that by knowing the structure of a protein one could predict its biological function. Today's research in molecular biology is more centered around the *organizational structure* of biological mechanisms. In this way, the ambiguity of the term 'structure' suits to uphold the research slogan, since it can also be applied in a broader sense here than just molecular shapes. The organization of biological systems is the domain of the relatively

new discipline systems biology.

The three-dimensional shape is often a detail that does not contribute to the understanding of a mechanism, but to the contrary would only confuse the mechanistic picture which requires a certain level of abstraction in order to be comprehensive.

But still, how exactly do we get from molecular structures and their (structured) activities to biochemical activities and biological functions? That there might not exist a straightforward mapping from molecular shapes to their biochemical and biological function had been anticipated in the early days of molecular biology:

> "It [molecular biology] is concerned particularly with the *forms* of biological molecules, and with the evolution, exploitation and ramification of these forms in the ascent to higher and higher levels of organization. Molecular biology is predominantly three-dimensional and structural—which does not mean, however, that it is merely a refinement of morphology. It must of necessity enquire at the same time into genesis and function." (Astbury, 1952, 3, original emphasis)

Taking up Francis Crick's remark that "folding is simply a function of the order of the amino acids" (Crick 1958, 144), Morange comments that it is "obviously not a *simple* function" (Morange, 2006, 522). And he observes a semantic change in the meaning of 'function':

> "For Francis Crick, function meant the application of simple rules and principles. For specialists today, function is the result of a complex evolution [...] This shift in the meaning of a word is more than anecdotal. It reflects an active ongoing transformation of biology [...] The mechanistic models of

> molecular biology are no longer considered sufficient to explain the structures and functions of organisms. They have to be complemented and allied with evolutionary explanations" (Morange, 2006, 522).

In order to explain biological phenomena, there is no determination relation that would allow us to track everything down to the chemical and physical properties of proteins, let alone the nucleotide sequences of DNA. Of course, all these issues are relevant to the topic of reduction:

> "if [...] regulatory networks turn out to be crucial to explaining development (and evolution [...]), the reductionist interpretation *may* be in trouble. If network-based explanations are ubiquitous, it is quite likely that what will often bear the explanatory weight in such explanations is the topology of the network rather than the specific entities of which it is composed. [...] How topological an explanation is becomes a matter of degree: the more an explanation depends on individual properties of a vertex, the closer an explanation comes to traditional reduction. The components matter more than the structure. Conversely, the more an explanation is independent of individual properties of a vertex, the less reductionist it becomes." (Sarkar, 2008, 68, original emphasis)

5 Conclusion

Both terms, 'structure' and 'function', are highly ambiguous. So is the widely used conjunct phrase of 'structure and function' that is ubiquitous in biology, as well as the

even stonger claim 'structure determines function'. Perhaps this is why it can be used in many different contexts and for many different explanatory aims in biology. Although providing a certain framework of generality, I argue that a clarification of these concepts is beneficial—for conceptual and philosophical considerations, as well as for the way biologists think about the grand schemes like the "central dogma". Ideally, such an account would also have practical implications and benefit current biological research.

To sum up the results of my analysis, in molecular biology's explanatory practice, 'structure' may refer to:

1. the sequence of polymers,
2. the three-dimensional shape of molecules (or their parts), and
3. the way biological entities are organized.

Of course, different aspects of this distinction play different roles in the explanatory practice with respect to molecular mechanisms. The detailed shape of the interacting molecules is neither necessary nor sufficient for understanding its activities (although correlations are valuable prediction tools before doing experiments in the lab).

The ambiguity of the term 'function' depends on whether the explanation aims at answering the question how a mechanism works or how it came to work that way. Even in the first case one has to distinguish between:

1. the biochemical activity of individual components, and
2. the biological role of network structures.

Whereas biochemical activities of proteins can often be successfully predicted by homology modeling from known molecular shapes, the biological role is rarely an

intrinsic property of an isolated molecule. Rather, the biological role is the mechanistic result of an interaction network of several dynamically interacting molecules.

By comparing the combinatorial possibilities of the different meanings of 'structure' and 'function', a determination relation does not hold between any of them. Instead, I propose a supervenience relation: between the three-dimensional shapes of protein domains and their biochemical activities, and between interaction networks and their biological role. According to my analysis, this is what molecular biologist mean when they say 'structure determines function'.

References

Alexander, P. A., Y. He, Y. Chen, J. Orban, and P. N. Bryan (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences 104*(29), 11963–11968. doi:10.1073/pnas.0700922104.

Allen, C. (2009). Teleological notions in biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009 ed.). `http://plato.stanford.edu/archives/win2009/entries/teleology-biology/`.

Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science 301*(5641), 1866–1867. doi:10.1126/science.1089072.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science 181*(4096), 223–230. doi:10.1126/science.181.4096.223.

Anfinsen, C. B., E. Haber, M. Sela, and F. H. White, Jr (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences 47*(9), 1309–1314.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. *Nature Genetics 25*, 25–29. doi:10.1038/75556.

Astbury, W. T. (1952). Adventures in molecular biology. In *The Harvey Lectures. Delivered under the auspices of the Harvey Society of New York. 1950–51*, pp. 3–44. Charles C Thomas.

Burian, R. M. (1996). Underappreciated pathways toward molecular genetics as illustrated by Jean Brachet's cytochemical embryology. In S. Sarkar (Ed.), *The Philosophy and History of Molecular Biology: New Perspectives*, pp. 67–85. Kluwer Academic Publishers.

Chouard, T. (2011). Breaking the protein rules. *Nature 471*, 151–153. doi:10.1038/471151a.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy 72*(20), 741–765. doi:10.2307/2024640.

Forman-Kay, J. D. and T. Mittag (2013). From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure 21*(9), 1492–1499. doi:10.1016/j.str.2013.08.001.

Green, S. (2017). Philosophy of systems and synthetic biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 ed.). `https://plato.stanford.edu/archives/sum2017/entries/systems-synthetic-biology/`.

Gutte, B. and R. B. Merrifield (1971). The synthesis of ribonuclease A. *Journal of Biological Chemistry 246*, 1922–1941.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray (1999). From molecular to modular cell biology. *Nature 402*(6761 Suppl.), C47–C52. doi:10.1038/35011540.

Hvidsten, T. R., A. Lægreid, A. Kryshtafovych, G. Andersson, K. Fidelis, and J. Komorowski (2009). A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLoS ONE 4*(7), e6266. doi:10.1371/journal.pone.0006266.

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, and H. Wyckoff (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature 181*(4610), 662–666. doi:10.1038/181662a0.

McLaughlin, B. and K. Bennett (2018). Supervenience. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). `https://plato.stanford.edu/archives/spr2018/entries/supervenience/`.

Millikan, R. G. (1989). An ambiguity in the notion "function". *Biology and Philosophy 4*, 172–176. doi:10.1007/BF00127747.

Mitchell, S. D. and A. M. Gronenborn (2017). After fifty years, why are protein X-ray

crystallographers still in business? *The British Journal for the Philosophy of Science 68*(31), 703–723. doi:10.1093/bjps/axv051.

Morange, M. (2006). The protein side of the central dogma: Permanence and change. *History and Philosophy of the Life Sciences 28*(4), 513–524.

Pal, D. and D. Eisenberg (2005). Inference of protein function from protein structure. *Structure 13*, 121–130. doi:10.1016/j.str.2004.10.015.

Rose, G. D. and T. P. Creamer (1994). Protein folding: Predicting predicting. *PROTEINS: Structure, Function, and Genetics 19*, 1–3.

Roy, A., A. Kucukural, and Y. Zhang (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols 5*(4), 725–738. doi:10.1038/nprot.2010.5.

Sadowski, M. and D. T. Jones (2009). The sequence–structure relationship and protein function prediction. *Current Opinion in Structural Biology 19*, 357–362. doi:10.1016/j.sbi.2009.03.008.

Sarkar, S. (2008). Genomics, proteomics, and beyond. In S. Sarkar and A. Plutynski (Eds.), *A Companion to the Philosophy of Biology*, pp. 58–73. Blackwell Publishing Ltd.

Shapiro, J. A. (2011). *Evolution: a view from the 21st century.* FT Press Science.

Wouters, A. G. (2003). Four notions of biological function. *Studies in History and Philosophy of Biological and Biomedical Sciences 34*(4), 633–668. doi:10.1016/j.shpsc.2003.09.006.

Wright, L. (1973). Functions. *The Philosophical Review 82*(2), 139–168. doi:10.2307/2183766.

Wright, P. E. and H. J. Dyson (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology 293*, 321–331. doi:10.1006/jmbi.1999.3110.

Zhou, M., J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. Sachs, and Y. Liu (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature 495*, 111–115. doi:10.1038/nature11833.

Is Peer Review a Good Idea?*

Remco Heesen[†‡] Liam Kofi Bright[§]

September 19, 2018

Abstract

Pre-publication peer review should be abolished. We consider the effects that such a change will have on the social structure of science, paying particular attention to the changed incentive structure and the likely effects on the behavior of individual scientists. We evaluate these changes from the perspective of epistemic consequentialism. We find that where the effects of abolishing pre-publication peer review can be evaluated with a reasonable level of confidence based on presently available evidence, they are either positive or neutral. We conclude that on present evidence abolishing peer review weakly dominates the status quo.

*Both authors contributed equally. Thanks to Justin Bruner, Adrian Currie, Cailin O'Connor, and Jan-Willem Romeijn for valuable comments. RH was supported by an Early Career Fellowship from the Leverhulme Trust and the Isaac Newton Trust. LKB was supported by NSF grant SES 1254291.

[†]Department of Philosophy, School of Humanities, University of Western Australia, Crawley, WA 6009, Australia. Email: remco.heesen@uwa.edu.au.

[‡]Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK.

[§]Department of Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: liamkbright@gmail.com.

1 Introduction

Peer review plays a central role in contemporary academic life. It sits at the critical juncture where scientific work is accepted for publication or rejected. This is particularly clear when the results of scientific work are communicated to non-scientists, e.g., by journalists. The question "Has this been peer reviewed?" is commonly asked, and a positive answer is frequently taken to be a necessary and sufficient condition for the results to be considered serious science.

Given these circumstances, one might expect peer review to be an important topic in the philosophy of science as well. Peer review should arguably play a more prominent role in the debate about demarcation criteria (what separates science from other human pursuits?), as it seems to be used in practice exactly to differentiate scientific knowledge from other claims to knowledge, at least by journalists. Yet as far as we know, social-procedural accounts of science, like the one found in Longino (1990), remain in the minority and usually do not place great emphasis on peer review in particular. Aside from this particular debate, there are normative questions about the proper epistemic role of peer review and more practical questions about the extent to which it manages to fulfill them, all of which should interest philosophers of science.

But there has been surprisingly little work on peer review by philosophers of science. Most of what exists has focused on the role of biases in peer review, see for example Saul (2013, §2.1), Lee et al. (2013), Jukola (2017), Katzav and Vaesen (2017), and Heesen (2018). We are not aware of any philosophical discussion of the strengths and weaknesses of peer review as such (the above examples presuppose its overall legitimacy by discussing its implementation). Some work along these lines does exist outside of philosophy, either in the form of opinion pieces (Gowers 2017) or occasionally full-length articles (Smith 2006). Such work tends to be vague about the normative standard against which peer review or its alternatives are to be

Seattle, WA; 1-4 November 2018 -301-

evaluated, something we aim to remedy in section 2.

Here we bring together the work of philosophers of science (especially social epistemologists of science) who have written about the strengths and weaknesses of various aspects of the social structure of science and empirical work about the effects of peer review. We argue that where philosophers of science have claimed the social structure of science works well, their arguments tend to rely on things other than peer review, and that where specific benefits have been claimed for peer review, empirical research has so far failed to bear these out. Comparing this to the downsides of peer review, most prominently the massive amount of time and resources tied up in it, we conclude that we might be better off abolishing peer review.

Some brief clarifications. Our target is pre-publication peer review, that is the review of a manuscript intended for publication, where publication is withheld until one or more editors deem the manuscript to have successfully passed peer review. We set aside other uses of peer review (e.g., of grant proposals or conference abstracts) and we explicitly leave room for post-publication peer review, where manuscripts are published before review. Because of this last point, some readers may think that our terminology ('abolishing pre-publication peer review') suggests a more dramatic change than what we actually advocate. We invite such readers to substitute in their preferred terminology. We should also clarify that we use 'science' in a broad sense to include the natural sciences, the social sciences, and the humanities.

The overall structure of our argument is as follows. We think there are a number of clear benefits to abolishing pre-publication peer review. In contrast, while various benefits of the existing system (downsides of abolishing peer review) have been suggested, we do not think there exist any that have clear empirical support. Insofar as empirical research exists, it is ambiguous in some cases, and speaks relatively clearly against the claimed benefit of the existing system in others. While we admit to a number of cases where the evidence is ambiguous or simply lacking (see especially section 5), we claim

that the present state of the evidence suggests that abolishing pre-publication peer review would lead to a Pareto improvement: each factor considered is either neutral or favors our proposal.

Our primary aim here is to evaluate the current system, but we believe that is only really possible by comparing it to an alternative. We are not claiming that the proposal we put forward is the best of all possible alternatives. It has been constructed to be a system which could constitute a Pareto improvement over the current system. Given that it has not actually been implemented yet, we cannot guarantee it would work as advertised or what empirical properties it would have. But in offering a relatively specific alternative, we hope to get people thinking about real change, which pointing out problems with the present system has so far failed to do.

Even despite this proviso, we realize that ours is a strong claim, and our proposal a large change to the social structure of science. It is therefore important to highlight that our central claim concerns the balance of presently available evidence. We are not further claiming that the matter is so conclusively settled as to render further research superfluous or wasteful. On the contrary, we think there are a number of points in our argument where the presently available evidence is severely limited, and we take the calls for further empirical research we make in those places to be just as important a part of the upshot of our paper as our positive proposal. We hope, therefore, that even a skeptical reader will read on; if not to be convinced of the need of abolishing pre-publication peer review, then at least to see where in our view their future research efforts should concentrate if they are to shore up pre-publication peer review's claims to good epistemic standing.

2 Setting the Stage

The purpose of peer review is usually construed in terms of quality control. For example, Katzav and Vaesen (2017, 6) write "The epistemic role of peer

Seattle, WA; 1-4 November 2018 -303-

review is assessing the quality of research", and this seems to be a common sentiment per, e.g., Eisenhart (2002, 241) and Jukola (2017, 125). But how well does peer review succeed in its purpose of quality control? The empirical evidence (reviewed below) is mixed at best. As one prominent critic puts it, "we have little evidence on the effectiveness of peer review, but we have considerable evidence on its defects" (Smith 2006, 179).

Peer review's limited effectiveness would perhaps not be a problem if it required little time and effort from scientists. But in fact the opposite is true. Going from a manuscript to a published paper involves many hours of reviewing work by the assigned peer reviewers and a significant time investment from the editor handling the submission. The editor and reviewers are all scientists themselves, so the epistemic opportunity cost of their reviewing work is significant: instead of reviewing, they could be doing more science.

Given these two facts—high (epistemic) costs and unclear benefits—we raise the question whether it might be better to abolish pre-publication peer review. In the following we provide our own survey and assessment of the evidence that bears on this question. Our conclusions are not sympathetic to peer review. However, we encourage any proponents of peer review to give their own assessment. We only ask that any benefits claimed for peer review are backed up by empirical research, and that they are epistemic benefits, i.e., we ask for empirical evidence that peer review makes for better science on science's own terms.

We take the status quo to be as follows. The vast majority of scientific work is shared through journal publications, and the vast majority of journals uses some form of pre-publication peer review. Ordinarily this means that an editor assigns one to three peers (scientists whose expertise intersects the topic of the submission), who provide a report and/or verdict on the submission's suitability for publication. The peer reviews feed into the final judgment: the submission is accepted or rejected with or without revisions.

Our proposal is to abolish pre-publication peer review. Scientists them-

selves will decide when their work is ready for sharing. When this happens, they publish their work online on something that looks like a preprint archive (although the term "preprint" would not be appropriate under our proposal). Authors can subsequently publish updated versions that reply to questions and comments from other scientists, which may have been provided publicly or privately. Most journals will probably cease to exist, but the business of those that continue will be to create curated collections of previously published articles. Their process for creating these collections will presumably still involve peer review, but now of the post-publication variety.

Our proposal is in line with how certain parts of mathematics and physics already work: uploading a paper to arXiv is considered publishing it for most purposes, with journal peer review and publication happening almost as an afterthought (Gowers 2017). Indeed, journal publication can function as something like a prize, accruing glory to the scientist who achieves it but doing little to actually help spread or diffuse the idea beyond calling attention to something that has already been made public elsewhere. We are not aware of any detailed comparative studies of the effects these changes have had in those fields, so we will not rest any significant part of our argument on this case. But for those who worry that science will immediately and irrevocably fall apart without peer review, we point out that this does not appear to have happened in the relevant parts of mathematics and physics.

In the remainder of this paper we break down the consequences of our proposal. Our strategy here is to focus on a large number (hopefully all) aspects of the social structure of science that will be affected. In particular, the reader may already have a particular objection against our proposal in mind. We encourage such a reader to skip ahead to the section where this objection is discussed before reading the rest of the paper.

For example, one reader may think that peer review as currently practiced is important because it forces scientists to read and review each other's work, and without peer review they will spend less time on such tasks. This

Seattle, WA; 1-4 November 2018 -305-

is discussed in section 3.2. Another reader may worry that without peer review and the journal publications that go with them it will be more difficult to evaluate scientists for hiring or promotion (section 3.5). Yet another reader may be concerned about losing peer review's ability to prevent work of little merit from being published, or at least to sort papers into journals by epistemic merit so scientists can easily find good work (section 4.1). A fourth reader might think peer review plays an important role in detecting fraud or other scientific malpractice (section 4.2). A fifth reader may think the guarantee provided to outsiders when something has been peer reviewed is an important reason to preserve the status quo (section 5.1). And a sixth reader may want to point out that anonymized peer review gives relatively unknown scientists a chance at an audience by publishing in a prestigious journal, whereas on our proposal perhaps only antecedently prominent scientists will have their work read and engaged with (section 5.2).

Other aspects of the social structure of science that will be considered: whether and when scientists share their work (section 3.1), how many papers are published by women or men (section 3.3), library resources (section 3.4), the power of editors as gatekeepers (section 3.6), science's susceptibility to fads and fashions (section 4.3), and ways to get credit for scientific work other than through journal publications (section 4.4). In each case we evaluate whether the net effects of our proposal on that aspect can be expected to be positive. To tip our hand: aspects where we will claim a benefit are gathered in section 3, aspects where we expect little or no change are in section 4, and aspects that we consider neutral due to a present lack of evidence are in section 5.

In making these evaluations, we commit to a kind of epistemic consequentialism (cf. Goldman 1999). One may think of what we are doing as roughly analogous to the utilitarian principle, where for each issue our yardstick is whether pre-publication peer review shall generate the greatest amount of knowledge produced in the least amount of time. More specifically, we con-

sider changes in the incentive structure and expected behaviour of scientists, as well as other changes that would result from abolishing pre-publication peer review. We evaluate these changes in terms of their expected effect on the ability of the scientific community to produce scientific knowledge in an efficient manner. Working out in detail what such an epistemic consequentialism would look like would be very complicated, and we do not attempt the task here. For most of the issues we consider, we think that the calculus is sufficiently clear that fine details do not matter. Where it is unclear (the issues discussed in section 5) we think this results from ignorance of empirical facts about the likely effect of policies, rather than conceptual unclarity in the evaluative metric. So we do not need to use our consequentialist yardstick to settle any difficult tradeoffs. All we need for our purposes is to make it clear that we are evaluating the peer review system by how well it does in incentivizing efficient knowledge production.

What do we mean by the incentive structure of science, mentioned in the previous paragraph? This addresses the motivations of scientists. Scientists are rewarded for their contributions with credit, i.e., with recognition from their peers as expressed through such things as awards, citations, and prestigious publications (Merton 1957, Hull 1988, Zollman 2018). Scientific careers are largely built on the reputations scientists acquire in this way (Latour and Woolgar 1986, chapter 5). As a result, scientists engage in behaviors that improve their chances of credit (Merton 1969, Dasgupta and David 1994, Zollman 2018).

While individual scientists may be motivated by credit to different degrees (curiosity, the thrill of discovery, and philanthropic goals are important motivations for many as well), the effect on careers means that credit-maximizing behavior is to some extent selected for. Thus we think it important to ensure that our proposal does not negatively affect the incentives currently in place for scientists to work effectively and efficiently.

3 Benefits of Abolishing Peer Review

3.1 Sharing Scientific Results

An important feature of (academic) science is that there is a norm of sharing one's findings with the scientific community. This has been referred to as the communist norm (Merton 1942). In recent surveys, scientists by and large confirm both the normative force of the communist norm and their actual compliance (Louis et al. 2002, Macfarlane and Cheng 2008, Anderson et al. 2010). This norm is epistemically beneficial to the scientific community, as it prevents scientists from needlessly duplicating each other's work.

Will abolishing peer review affect this practice? In order to answer this question, we need to know what motivates scientists to comply with the communist norm, that is to share their work. On the one hand there is the feeling that they ought to share generated by the existence of the norm itself. There is no reason to expect this to be changed by abolishing peer review.

On the other hand there is the motivation generated by the desire for credit. According to the priority rule, the first scientist to publish a particular discovery gets the credit for it (Merton 1957, Dasgupta and David 1994, Strevens 2003). So a scientist who wants to get credit for her discoveries has an incentive to publish them as quickly as possible, in order to maximize her chances of being first. Recent work suggests that this applies even in the case of smaller, intermediate discoveries (Boyer 2014, Strevens 2017, Heesen 2017b). All of this helps motivate scientists to share their work.

If peer review were to be abolished, the communist norm and the priority rule would still be in effect, so scientists would still be motivated to share their work as quickly as possible. However, the following change would occur.

In the absence of pre-publication peer review, scientists would be able to share their discoveries more quickly. In the current system, peer review can hold up publication for significant amounts of time, especially in the case of fields with high rejection rates or long turnaround times. During this time,

other scientists cannot build on the work and may spend their time needlessly duplicating the work. Cutting out this lag by letting scientists publish their own work when they think it is ready will speed up scientific progress. While being faster is not always better (it may increase the risk of error, cf. Heesen 2017c), in this case delays in publication are reduced without any reduction in the time spent on the scientific work itself.

To some extent this already occurs. Scientists, especially well-connected scientists, already share preprints that make the community aware of their work in advance of publication. For people who regularly do this, practically speaking little would change upon adopting the system we advocate. However, our proposal turns pre-journal-publication dispersal of work from a privilege of a well-connected few into the norm for everyone.

On this point, then, abolishing peer review is a net positive, as scientists will still be incentivized to share their work as soon as possible, but the delays associated with pre-publication peer review are removed.

3.2 Time Allocation

The current system restricts the way scientists are allowed to spend their time. For each paper submitted to a journal, a number of scientists are conscripted into reviewing it, and at least one editor has to spend time on that paper as well.

On our proposal, scientists would be free to choose how much of their time to spend reading and reviewing others' work as compared to other scientific activities. Some scientists would spend less time reviewing, some scientists would spend more, and some would spend exactly as much as under the current system.

For scientists in the latter category our proposal makes no difference, while for scientists in the other two categories our proposal represents a net improvement of how they spend their time, at least in their own judgments. We think people are the best judges of how to use their own time and labor.

We thus trust scientists' decisions in these regards, and welcome changes that would render many scientists' choices about how to allocate their own labor independent of the preferences of the relatively small number of editorial gatekeepers.

So we assume that scientists are well-placed to judge how best to use their own abilities to meet the community's epistemic needs. We claim, moreover, that the reward structure of science is set up so as to make it in their interest to do so: the credit economy incentivizes scientists to spend their time on pursuits the epistemic value of which will be recognized by the community (Zollman 2018). Hence freeing up the way scientists allocate their time leads to net epistemic benefits to the scientific community.

One might object that journals perform a useful epistemic sorting role, telling scientists what is worth spending their time on. We will address these concerns in section 4.1.

One might think that this would lead scientists to spend significantly less time reading and reviewing others' work. If this is right, we still think it would be an overall improvement for the reasons mentioned above. But we also want to point out that this is not as obvious a consequence as it may seem. Here are two reasons to expect scientists to spend as much time or more reading and reviewing on our proposal. First, for many scientists reading and reviewing are intrinsically valuable and can help their own research. Second, the current system provides no particular incentive to read and review either: scientists agree to review only because they independently want to or because they feel an obligation to the research community. While no one scientist is conscripted, at the group level editors are going to keep going until they find someone. This can amount to picking whomever is most weak-willed or under some extra-epistemic social pressure. It is not obvious that this way of deciding who does the reviewing has much to recommend it. Any rewards that exist for reviewing will still exist on our proposal, and may be amplified by the possibility of making post-publication reviews public.

3.3 Gender Skew in Publications

Male scientists publish more, on average, than female scientists, a phenomenon known as the productivity puzzle or productivity gap (Zuckerman and Cole 1975, Valian 1999, Prpić 2002, Etzkowitz et al. 2008). Several explanations have been suggested, none of which are entirely satisfactory (see especially Etzkowitz et al. 2008, 409–412). Two of these explanations that are relevant to our concerns here are the direct effects of gender bias and the indirect effects of the expectation of gender bias.

There is some evidence of gender bias in peer review, although this is not unambiguous (see Lee et al. 2013, 7–8, Lee 2016, and references therein). Insofar as there is gender bias—in the sense of women's work being judged more negatively by peer reviewers—abolishing peer review will remove this and help level the playing field for men and women. We expect positive epistemic consequences from the removal of these arbitrarily different standards.

While the evidence of gender bias in peer review is not entirely clear-cut, there is good evidence that women *expect* to face gender bias in peer review (see Lee 2016, Bright 2017b, Hengel 2018, and references therein). In an effort to overcome this perceived bias, women tend to hold their own work to higher standards. Hengel (2018) provides evidence that women spend more time correcting stylistic aspects of their paper during peer review, presumably due to higher expectations of scrutiny on such apparently superficial elements of their work. On the plausible assumption that if women have higher standards for each paper they will produce fewer papers overall, this means that the mere expectation of gender bias can contribute to the productivity gap.

After abolishing peer review both women and men will hold their work primarily to their own individual standards of quality, and secondarily to their expectations of the response of the entire scientific community, but not to their expectations of the opinion of a small arbitrary group of gatekeepers. We do not know whether this will lead the women to behave more like the men (producing more papers) or the men to behave more like the women

(holding individual papers to a higher standard of quality). However, in line with our view above that scientists are well-placed to judge how best to spend their own time, we take it that any resulting change in behavior will be a net epistemic positive.

3.4 Library Resources

Journal subscription fees currently take up a large amount of library resources (RIN 2008, Van Noorden 2013). To summarize some key figures from the 2008 report: research libraries in the UK spent between £208,000 and £1,386,000 on journal subscriptions annually (and that was a decade ago, with subscriptions having risen substantially since). The cost for publishing and distributing a paper was estimated to be about £4,000, or about £6.4 billion per year in total. Savings from moving to author-paid open access were estimated at £561 million, about half of which would directly benefit libraries.

On our proposal, this is replaced by the cost of maintaining one or more online archives of scientific publications. The example of existing large preprint archives like arXiv and bioRxiv suggests that maintaining such archives can be done at a fraction of the cost currently spent on journal subscription fees. As a rough guideline, Van Noorden (2013) estimates maintenance costs of arXiv at just $10 per article. So our proposal involves significant savings on library resources, which could be used to expand collections, retain more or better trained staff, or other purposes that would be of epistemic benefit to the scientific community.

Two additional effects should be considered in relation to this. First, the fact that the online archive will be open access means that scientific publications will be available to everyone, not just to those with a library subscription or some other form of access to for-profit scientific journals.

Second, the fact that any value added by for-profit journals would be taken away. The two tasks currently carried out by journals that could

plausibly be supposed to add value to scientific publications are peer review and copy-editing (Van Noorden 2013). It is the purpose of all other sections of this paper to argue that peer review does not in fact (provably) add value, so we set that aside. This leaves copy-editing. We propose that libraries use some of the funds freed up from journal subscriptions to employ some copy-editors. Each university library would make copy-editing services available to the scientists employed at that university. We contend that, after paying for the maintenance of an online archive and a team of copy-editors, under our proposal libraries would still end up with more resources for other pursuits than under the current system.

We note that this particular advantage of our proposal is a bit more historically contingent than the others. There seems to be no particular reason why pre-publication peer review has to be implemented through for-profit journals, and if the open access movement has its way we might be able to free up these library resources without abolishing pre-publication peer review. But our proposal also achieves this goal, and so we count it as an advantage relative to the system as it is currently actually implemented.

3.5 Scientific Careers

The 'publish or perish' culture in science has been widely noted (e.g., Fanelli 2010). Universities judge the research productivity of scientists through their publications in (peer reviewed) journals, with some focusing more on 'quantity' (counting publications) and others on 'quality' (publishing in prestigious journals). Scientific journals and the system of pre-publication peer review thus play an important role in shaping scientific careers. What will become of this if peer review is abolished?

We note first that the 'publish or perish' culture is a subset of a larger system which we discussed above: the credit economy. Publishing in a journal is one way to receive credit for one's work, but there are others, most prominently citations and awards. Scientific careers depend on all of these,

with different institutions weighting quantity of publications, quality of publications, citation metrics, and awards and other honors differently.

Any of these types of credit represents some kind of recognition of the scholarly contributions of the scientist by her peers. But here we distinguish two types of credit, which we will call short-run credit and long-run credit. Getting a paper through peer review yields a certain amount of credit: more for more prestigious journals, less for less prestigious ones. But this is short-run credit in the following sense. The editor and the peer reviewers judge the technical adequacy and the potential impact of the paper, shortly after it is written. Their judgment is essentially a prediction of how much uptake the paper is likely to receive in the scientific community.

In contrast, citations (as well as awards, prizes, inclusion in anthologies or textbooks, etc.) represent long-run credit. They *are* the uptake the paper receives in the scientific community. Long-run credit is both a more considered opinion of the scientific importance of the paper and a more democratic one (citations can be made by anyone, and awards usually reflect a consensus in the scientific community, whereas peer review is normally done by up to three individuals). So long-run credit reflects a more direct and better estimate of the real epistemic value of a contribution to science.

So what would the effect of our proposal be? For better or worse, our proposal does not make it impossible for universities to use metrics to judge research productivity. While journal rankings and impact factors would disappear, citation metrics for individual scientists and papers would still be available. This may mean that universities stop judging their scientists based on the impact factors of the journals they publish in and start judging them on the actual citation impact of their papers. More generally, our proposal will decrease or remove the role of short-run credit in shaping career outcomes and increase the role of long-run credit, which we take to be a better measure of scientific importance. So we think this is an improvement on the status quo.

What about junior hires and related career decisions, where long-run credit may be absent or minimal? If abolishing peer review means completely getting rid of journals and the associated prestige rankings, this robs hiring departments of some information regarding the scientific importance of candidates' work. If this means those on the hiring side need to read and form an opinion of candidates' work for themselves, we do not think that is a bad thing. This would of course take time, but if journals and peer review are completely abolished, that just means the time spent reviewing the paper is transferred to the people considering hiring the scientist, which again, we do not think is a bad thing. In fact, since very few academics are on a hiring committee year after year, whereas referee requests are a constant feature while one is in the community, we think that even this added burden when hiring might still be a net time-saver for academics.

But it does not have to be that way. We never said journals and peer review have to be completely abolished—our proposal in section 2 explicitly suggests journal issues may still appear, but as curated collections of articles based on post-publication peer review. So short-run credit based on journal prestige need not disappear. It need not even be slower as there is no particular reason post-publication peer review needs to take longer than pre-publication peer review. But there is the added advantage that the paper is already published while it undergoes peer review, so the wider community outside the assigned reviewers also has a chance to respond before it is included in a journal.

3.6 The Power of Gatekeepers

The discussion immediately above touched on another effect, one that we think is worth bringing out as a benefit of our proposal in its own right. As mentioned our proposal suggests that in evaluating the importance of scientific work we decrease our reliance on short-run credit (journal prestige), with a corresponding increase in long-run credit (citations, among other things).

This means that the overall credit associated with a particular paper depends less on the judgments made by an editor and a small number of reviewers, and more on its actual uptake in the larger scientific community.

Editors in particular currently play a large role in determining which scientific work is worthy of attention, as they are a relatively small group of people with a deciding vote in the peer review process of a large number of papers. They are often referred to as gatekeepers for this reason (Crane 1967). Our proposal entails significantly decreasing both the prevalence and importance of this role. By replacing some of this importance with long-run credit, which comes from the scientific community as a whole, it makes the evaluation of scientific work a more democratic process. Not only is there some reason to think that democratic evaluation of scientific claims is more in line with general communal norms accepted within science (Bright et al. 2018), but general arguments from democratic theory and social epistemology of science give epistemic reason to welcome the increased independence of judgment and evaluation this would introduce (List and Goodin 2001, Heesen et al. forthcoming, Perović et al. 2016, 103–104).

4 Where Peer Review Makes No Difference

In this section we consider a number of aspects of the scientific incentive structure for which we think a case can be made that abolishing peer review will leave them basically unaffected. This serves partially to forestall objections to our proposal that we anticipate from defenders of the peer review system, and partially to avoid overstating our case—in some of what follows we argue that abolishing peer review will likely have no effect in cases where one might have expected it to be beneficial.

4.1 Epistemic Sorting

Given the stated purpose of peer review mentioned in section 2 the first and most apparent disadvantage of our proposal is that it would remove the epistemic filter on what enters into the scientific literature. One might worry that the scientific community would lose the ability to maintain its own epistemic standards, and thus the general quality of scientific research would be reduced. We argue here that despite the intuitive support this idea might have, the present state of the literature on scientific peer review does not support it.

Separate out two kind of epistemic standards one may hope that the peer review system maintains. First, that peer review allows us to identify especially meritorious work and place it in high profile journals, while ensuring that especially shoddy work is kept from being published. Call this the 'epistemic sorting' function of peer review. Second, that peer review allows for the early detection of fraudulent work or work that otherwise involves research misconduct. Call this the 'malpractice detection' function of peer review. We deal with each of these in turn.

Let us step back and ask why, from the point of view of epistemic consequentialism, one would want peer review to do any sort of epistemic sorting. We take the answer to be that epistemic sorting helps scientists fruitfully direct their time and energy by selecting the best work and bringing it to scientists' attention through publication in journals. They read and respond to that which is most likely to help them advance knowledge in their field.

How could peer review achieve this? One might hope that peer review functions by keeping bad manuscripts out of the published literature and letting good manuscripts in. This, however, is a non-starter. There are far too many journals publishing far too many things, with standards of publication varying far too wildly between them, for the sheer fact of having passed peer review somewhere to be all that informative as to the quality of a manuscript.

Instead, if peer review is to serve anything like this purpose it must be because reviewers are able (even if imperfectly) to discern the relative degree of scientific merit of a work, and sort it into an appropriately prestigious journal. Epistemic sorting happens not via the binary act of granting or withholding publication, but rather through sorting manuscripts into journals located on a prestige hierarchy that tracks scientific merit.

A necessary condition for epistemic sorting to work as advertised is that reviewers be reliable guides to the merit of the scientific work they review. Our first critique is that this necessary condition does not seem to be met. Investigation into reviewing practices has not generally found much inter-reviewer reliability in their evaluations (Peters and Ceci 1982, Ernst et al. 1993, Lee et al. 2013, 5–6). What this means is that one generally cannot predict what one reviewer will think of a manuscript by seeing what another reviewer thought. If there was some underlying epistemic merit scientists were accurately (even if falteringly) discerning by means of their reviews, one would expect there to be correlations in reviewers evaluations. However, this is not what we find. Indeed, one study of a top medical journal even found that "reviewers. . . agreed on the disposition of manuscripts at a rate barely exceeding what would be expected by chance" (Kravitz et al. 2010, 3). Findings like these are typical in the literature that looks at inter-reviewer reliability (for a review of the literature see Bornmann 2011, 207). The available evidence does not provide much support for the idea that pre-publication peer review detects the presence of some underlying quality.

Our second critique of the epistemic sorting idea speaks more directly to the ideal it tracks. We are not persuaded that the best way to direct scientists' attention is to continually alert them to the best pieces of individual work, and have them proportion their attention according to position on a prestige hierarchy. We take it the intuition behind this is a broadly meritocratic one. This intuition has been challenged by some modeling work (Zollman 2009). While Zollman retained some role for peer review, his model

still found that striving to select the best work for publication is not necessarily best from the perspective of an epistemic community; his model favored a greater degree of randomization.

We do not wish to rest our case on the results of one model which in any case does not fully align with our argument, but it highlights that the ideal of meritocracy stands in need of more defense than it is typically given. We take it that scientists most fruitfully direct their attention to that package of previous work and results which, when combined, provides them with the sort of information and perspectives they need to best advance their own epistemically valuable projects. It is a presently undefended assumption that this package of work should be composed of works which are themselves individually the most meritorious work, or that paying attention to the prestige hierarchy of journals and proportioning one's attention accordingly will be useful in constructing such a package. Hence, even if it did turn out that the peer review system could sort according to scientific merit, it is an underappreciated but important fact that this is not the end of the argument. Further defense of the purpose of this kind of epistemic sorting is needed from the point of view of epistemic consequentialism.

Before moving on we note a potential objection. Even if one did not think that peer review was detecting some underlying quality or interestingness, one might think that the process of feedback and revision which forms part of the peer review system would be beneficial to the epistemic quality of the scientific literature. In this way epistemic sorting may have a positive epistemic effect even if it fails in its primary task.

However, this returns us to the points regarding gatekeepers and time allocation from section 3. We are not opposed to scientists reading each other's work, offering feedback, and updating their work in light of that. This can indeed lead to improvements (Bornmann 2011, 203), though in this context it is worth noting the results of an experiment in the biomedical sciences, which found that attempting to attach the allure of greater prestige

to more epistemically high caliber work did little to actually improve the quality of published literature (Lee 2013). Fully interpreting these results would require discussion of the measures of quality used in such literature. We do not intend to do that here, since we do not intend to dispute the point that it is desirable for scientists to give feedback and respond to it.

We would expect this sort of peer-to-peer feedback to continue under a system without pre-publication peer review. Curiosity, informal networking, collegial responsibilities, and the credit incentives to engage with others' work and make use of new knowledge before others do; these would all be retained even without pre-publication peer review. What would be eliminated is the assignment of reviewing duties to papers that scientists did not independently decide were worth their time and attention, and the necessity of giving uptake to criticism (in order to publish) independently of an author's own assessment of the value of that feedback.

We thus conclude that, from the point of view of epistemic consequentialism, there is presently little reason to believe that a loss of the epistemic sorting function of pre-publication peer review would be a loss to science. Inclusion in the literature does not do much to vouch for the quality of a paper; the evidence does not favor the hypothesis that reviewers are selecting for some latent epistemic quality in order to sort into appropriate journals; and the ideal underlying the claimed benefits of epistemic sorting is dubious. While peer reviewers do give potentially valuable feedback, there is no particular reason to think that changes in how scientists decide to spend their time would make things worse in this regard, and (per our arguments in section 3) some reason to think that they would make things better.

4.2 Malpractice Detection

The other way peer review might uphold epistemic standards is through malpractice detection. However, once again, the literature does not support this. A number of prominent cases of fraudulent research managed to sail

through peer review. Upon investigation into the behavior of those involved it was found there was no reason to think that peer reviewers or editors were especially negligent in their duties (Grant 2002, 3). Peer reviewers report unwillingness to challenge something as fraudulent even where they have some suspicion that this is so, and avoid the charge (Francis 1989, 11–12). A criminologist who looked into fraudulent behavior in science reported that "virtually no fraudulent procedures have been detected by referees because reading a paper is neither a replication nor a lie-detecting device" (Ben-Yehuda 1986, 6). A more recent survey of the evidence found, at the least, no consistent pattern in journals' self-reported ability to detect and weed out fraudulent results (Anderson et al. 2013, 235).

Even if the prospect of peer review puts some people off committing fraud, the fact that it is so unreliable at detecting fraud suggests that this is a very fragile deterrence system indeed. Even this psychological deterrence would be rapidly undermined by more adventurous souls, or those pushed by desperation, since many would quickly learn that pre-publication peer review is a paper tiger.

Conversely, there are various ways for malpractice detection to operate in the absence of peer review. These include motive modification (Nosek et al. 2012, Bright 2017a), encouraging post-publication replication and scrutiny (Bruner 2013, Romero 2017), and the sterner inculcation of the norms of science coupled with greater expectation of oversight among coworkers (Braxton 1990). All of these methods of deterring fraud or meliorating its effects would still be available under our proposal.

What evidence we now have gives little reason to suppose that abolishing pre-publication peer review is any great loss to malpractice detection. Thus in this regard our proposal would make no great difference to the epistemic health of science. Combining this with the discussion of epistemic sorting, we conclude there is presently no reason to believe pre-publication peer review is adding much value to science by upholding epistemic standards.

4.3 Herding Behavior

Where above we argued that pre-publication peer review is not making a positive difference often claimed for it, in this section we downplay a potential benefit of our proposal. A consistent worry about scientific behavior is that it is subject to fads or, in any case, some sort of undesirable herding behavior (see, e.g. Chargaff 1976, Abrahamson 2009, Strevens 2013). A natural thought is that pre-publication peer review encourages this, since by its nature it means that to get new ideas out there one must convince one's peers that the work is impressive and interesting. It has thus been claimed that pre-publication peer review encourages unambitious within-paradigm work that unduly limits the range of scientific activity (Francis 1989, 12). Reducing the incentive to herd might thus be claimed as a potential benefit of our proposal. However, we are not convinced that it is pre-publication peer review that is doing the harmful work here.

As mentioned above, our proposal eliminates or significantly reduces the importance of short-run credit, the credit that accrues to one in virtue of publishing in a (more or less prestigious) scientific journal. Long-run credit, on the other hand, is left untouched. Under any sort of credit system, a scientist needs to do work that the community will pay attention to, build upon, and recognize her for. The mere fact that (she believes that) her peers are interested in a topic and liable to respond to it is thus still positive reason to adopt a topic. This is true even if the scientist would not judge that topic to be the best use of her time if she were (hypothetically) free from the social pressures and constraints of the scientific credit system.

The best that could be said about our proposal in this regard is that scientists would not specifically have to pass a jury of peers before getting their work out there. But given that we anticipate continued competition for the attention of scientific coworkers, it is hard to say what the net effect in encouraging more experimental or less conformist scientific work would be.

Whatever conformist effects the credit incentive has (see also the discus-

sion immediately below) do not depend on whether it is short- or long-run credit one seeks. The conformism comes from the fact that credit incentives focus scientists' attention on the predicted reaction of their fellow scientists to their work. Pre-publication peer review might make this fact especially salient by bringing manuscripts before a jury of peers before they may be entered into the literature. But even without pre-publication peer review the credit-seeking scientist must be focused on her peers' opinions. So there is no particular reason to think that removing the pre-publication scrutiny of manuscripts will free scientists from their own anticipations of the fads and fashions of their day.

4.4 Long-Run Credit

We end this section by noting that many of the effects of the credit economy of science studied by social epistemologists really concern long-run credit rather than the short-run credit affected by retaining or eliminating pre-publication peer review. This point is not restricted to herding behavior.

For instance, social epistemologists have studied both the incentive to collaborate, and various iniquities that can arise when scientists do not start with equal power when deciding who shall do what work and how they shall be credited (Harding 1995, Boyer-Kassem and Imbert 2015, Bruner and O'Connor 2017, O'Connor and Bruner forthcoming). Whether or not manuscripts would have to pass pre-publication peer review in order to enter the scientific literature, there would still be benefits in the long run to collaboration, and (alas) there would still be social inequalities that allow for iniquities to manifest in the scientific prestige hierarchy.

For another example, social epistemologists have studied the ways in which the credit incentive encourages different strategies for developing a research profile or molding one's scientific personality to be more or less risk-taking (Weisberg and Muldoon 2009, Alexander et al. 2015, Thoma 2015). Once again, pre-publication peer review plays no particular role in the analy-

sis. The incentives to differentiate oneself from one's peers (without straying too far from the beaten path) and to mold one's personality accordingly exist independently of pre-publication peer review.

Two especially influential streams of work in the social epistemology of science have been the study of the division of cognitive labor (Kitcher 1990, Strevens 2003), and the role of credit in providing a spur to work in situations with a risk of under-production (Dasgupta and David 1994, Stephan 1996). These two streams have directed the focus of the field, and have formed some of the chief defenses of the credit economy of science as it now stands (but see Zollman 2018, for a more critical take).

We mention them here because pre-publication peer review or short-run credit again plays no particular role in the analyses offered by these papers. What drives their results is scientists' expectation that genuine scientific achievement will be recognized with credit. As we have argued above, it is long-run credit that best tracks genuine scientific achievement, and so it is long-run rather than short-run credit that grounds scientists' expectation in this regard. So in social epistemologists' most prominent defenses of the credit economy of science, long-run credit (while not named such) is the mechanism underlying the claimed epistemic benefits of the credit economy.

5 Difficulties For Our Proposal

We have discussed some benefits that would predictably accrue from abolishing peer review and some ways in which its apparent benefits are either under-evidenced or better attributed to the effects of long-run credit, which our proposal leaves untouched. We now discuss some cases which we take to be more problematic for our proposal—but by this point we hope to have at least convinced the reader that pre-publication peer review rests on shakier theoretical grounds than its widespread acceptance may lead one to suppose.

5.1 A Guarantee For Outsiders

One purpose pre-publication peer review serves is providing a guarantee to interested but non-expert parties. Science journalists, policy makers, scientists from outside the field the manuscript is aimed at, or interested non-scientists can take the fact that something has passed peer review as a stamp of approval from the field. At a minimum, peer review guarantees that outsiders are focusing on work that has convinced at least one relatively disinterested expert that the manuscript is worthy of public viewing. Given that there are real dangers to irresponsible science journalism or public action that is seen to be based on science that is not itself trustworthy (Bright 2018, §4), and that it is hard for non-experts to make the relevant judgment calls themselves, having a social mechanism to provide this kind of guarantee for outsiders is useful.

It is difficult to predict in advance what norms would come to exist for science journalists in the absence of pre-publication peer review. We thus first and foremost call for empirical research on this issue, possibly by studying what has happened in parts of mathematics and physics that already operate broadly along the lines we suggest (Gowers 2017).

However, against the presumption that things would be worse, we have two points to make. As the recent replication crisis has made clear, the value of peer review as a stamp of approval should not be overstated. There are reasons to doubt that peer review reliably succeeds in filtering out false results. We give three of them. First, peer reviewers face difficulties in actually assessing manuscripts—and just about anything can pass peer review eventually—as discussed under the heading of 'epistemic sorting' in section 4.1. Second, there are problems with the standards we presently use to evaluate manuscripts, in particular with the infamous threshold for statistical significance used in many fields (Ioannidis 2005, Benjamin et al. 2018). And third, deeper features of the incentive structure of science make replicability problems endemic (Smaldino and McElreath 2016, Heesen 2017c). Using

peer review as a stamp of approval may just be generating expert overconfidence (Angner 2006), without the epistemic benefits of greater reliability that would back this confidence up.

For the second part of our reply, recall that it is only pre-publication peer review that we seek to eliminate. We do not object to post-publication peer review resulting in papers being selected for inclusion in journals which mark the community's approval of such work, ideally after due and broad-based evaluation. If some such system were implemented then outsiders could use inclusion in such a journal as their marker of whether work is soundly grounded in the relevant science.

If such a stamp of approval from a journal or other communally recognized institution only comes a number of months or years after something is first published then we would expect it to represent a more well-considered judgment. Note that this would not necessarily slow the diffusion of knowledge as under the present system the same paper would have spent time hidden from view going through pre-publication peer review. The end result might not even be all that different from what happens in the present system, except that post-publication peer review would take into account more of the response or uptake from the wider scientific community. Thus it would more closely approximate the considered judgment of the community, as ultimately reflected in the long-run credit accorded to the paper.

5.2 A Runaway Matthew Effect

The second problem we are less confident we can deal with is that of exacerbating the Matthew effect. This is the phenomenon, first identified by Merton (1968), of antecedently more famous authors being credited more for work done simultaneously or collaboratively, even if the relative size or skill of their contribution does not warrant a larger share of the reward. Arguably the present system helps put a damper on the Matthew effect, allowing a junior or less prestigious author to secure attention for their work by publishing

in a high profile journal. Without such a mechanism to grab the attention of the field, perhaps scientists would just decide what to pay attention to based on their prior knowledge of the author or recommendation from others. This would strengthen the effects of networks of patronage and prestige bias favoring fancy universities. Thus squandering valuable opportunities to learn from those who were not initially lucky in securing a prestigious position or patronage from the already established.

While some have defended the Matthew effect (Strevens 2006), we will not go that route in defending our proposal for two reasons. First, the Matthew effect can perpetuate iniquities that themselves harm the generation and dissemination of knowledge (Bruner and O'Connor 2017). Second, even if it could be justified at the level of individual publications, its long-term effects are epistemically harmful. The scientific community allocates the resources necessary for future work on the basis of its recognition of past performance. So if there is excess reward for some and unfair passing over of others at the present stage of inquiry, this will ramify through to future rounds of inquiry, misallocating resources to people whose accomplishments do not fully justify their renown (Heesen 2017a). Hence on grounds of epistemic consequentialism we take seriously the problem of a runaway Matthew effect.

As mentioned, due to the pressures of credit-seeking and their own curiosity, scientists would still have incentive to read others' work and adapt it to suit their own projects. There is always a chance that valuable knowledge may be gathered from the work of one who has been ignored, which could provide an innovative edge. To some extent this creates opportunities for arbitrage: if the Matthew effect ever became especially severe there would be a credit incentive to specialize in seeking out the work of scientists who are not getting much attention. The lesson here is that the Matthew effect can only ever be so severe, before the credit incentive starts providing counter-veiling motivations.

However, this does not fully solve our problem. Moreover, so long as

resource allocation is tied to recognition of past performance the differences in recognition generated by the Matthew effect can and often do become self-fulfilling prophesies, as those with more gain the resources to do better in the future, and those without are starved of the resources necessary to show their worth.

It is not clear where to go from here. From the above it may seem like a solution would be to pair our proposal with a call to loosen the connection between recognition of a scientist's greatness based on their past performance and resource allocation. Indeed, this may well be independently motivated (Avin forthcoming, Heesen 2017a, §6). However, even short of this far-reaching change, we feel at present that this matter deserves more study rather than any definitive course of action.

Our present thought is that this is a very speculative objection, and there is no empirical evidence to back up the claim that eliminating pre-publication peer review will have dire consequences in this regard. In particular, while the present system may (rarely) allow a relative outsider to make a big splash, the common accusation of prestige bias in peer review (Lee et al. 2013, 7) suggests that on the whole pre-publication peer review may contribute to the Matthew effect rather than curtailing it.

More specifically, the Matthew effect can be made worse by peer review when anonymity breaks down in ways that systematically favor antecedently famous scientists. If this gives famous scientists more opportunities to publish papers, then our system may provide welcome relief, since it allows more people to get their papers out there. Hence whether our proposal makes the Matthew effect worse or better depends on whether the stronger influence would be who gets into the conversation (for which pre-publication peer review can exacerbate the Matthew effect), or who gets listened to once the conversation has begun (for which our proposal looks more problematic). Presently we cannot say which is the more significant effect. So, while we grant that a runaway Matthew effect may occur under our system, we prefer

to stress that at this point it is just not known whether the Matthew effect will be worse with or without pre-publication peer review.

What we propose is a large change, involving freeing up a lot of time and opening it up to more self-direction on the part of scientists, and it is not clear what sort of institutional changes it would be paired with. With more study of epistemic mechanisms designed especially to promote the work of junior or less prestigious scientists there might be found some way of surmounting the problem of a runaway Matthew effect, should it arise. Ultimately, only empirical evidence can settle these questions. Given the clear benefits and the unclear downsides of our proposal, we hope at minimum to inspire a more experimental attitude towards peer review.

6 Conclusion

Pre-publication peer review is an enormous sink of scientists' time, effort, and resources. Adopting the perspective of epistemic consequentialism and reviewing the literature on the philosophy, sociology, and social epistemology of science, we have argued that we can be confident that there would be benefits from eliminating this system, but have no strong reasons to think there will be disadvantages. There is hence a kind of weak dominance or Pareto argument in favor of our proposal.

To simplify things, imagine forming a decision matrix, with rows corresponding to 'Keeping pre-publication peer review' and 'Eliminating pre-publication peer review'. The columns would each be labeled with an issue studied by science scholars which we have surveyed here: gender bias in the literature, speed of dissemination of knowledge, efficient allocation of scientists' time and attention, etc. For each column, if there is a clear reason to think that either keeping or eliminating pre-publication scientific peer review does better according to the standards of epistemic consequentialism, place a 1 in the row of that option, and a 0 in the other. If there is no reason to

favor either according to present evidence, put a 0 in both rows.

Our present argument could then be summarized with: as it stands, the only 1s in such a table would appear in the row for eliminating pre-publication peer review. We thus advocate eliminating pre-publication peer review. Journals could still exist as a forum for recognizing and promoting work that the community as a whole perceives as especially meritorious and wishes to recommend to outsiders. Scientists would still have every reason to read, respond to, and consider the work of their peers; pre-publication peer review is not the primary drive behind either the intellect's curiosity or the will's desire for recognition, and either of those suffice to motivate such behaviors.

The overall moral to be drawn mirrors that of our invocation of the importance of long-run over short-run credit. The best guarantor of the long run epistemic health of science is science: the organic engagement with each others' ideas and work that arises from scientists deciding for themselves how to allocate their cognitive labor, and doing the hard work of replicating and considering from new angles those ideas that have been opened up to the scrutiny of the community. All this would continue without pre-publication peer review, and the best you can say for the system that currently uses up so much of our time and resources is that it often fails to get in the way.

References

Eric Abrahamson. Necessary conditions for the study of fads and fashions in science. *Scandinavian Journal of Management*, 25(2):235–239, 2009. doi: 10.1016/j.scaman.2009.03.005. URL `http://dx.doi.org/10.1016/j.scaman.2009.03.005`.

Jason McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. Epistemic landscapes, optimal search, and the division of cognitive

labor. *Philosophy of Science*, 82(3):424–453, 2015. doi: 10.1086/681766. URL http://dx.doi.org/10.1086/681766.

Melissa S. Anderson, Emily A. Ronning, Raymond De Vries, and Brian C. Martinson. Extending the Mertonian norms: Scientists' subscription to norms of research. *The Journal of Higher Education*, 81(3):366–393, 2010. ISSN 1538-4640. doi: 10.1353/jhe.0.0095. URL https://muse.jhu.edu/journals/journal_of_higher_education/v081/81.3.anderson.html.

Melissa S. Anderson, Marta A. Shaw, Nicholas H. Steneck, Erin Konkle, and Takehito Kamata. Research integrity and misconduct in the academic profession. In Michael B. Paulsen, editor, *Higher Education: Handbook of Theory and Research*, volume 28, chapter 5, pages 217–261. Springer, Dordrecht, 2013. doi: 10.1007/978-94-007-5836-0_5. URL http://dx.doi.org/10.1007/978-94-007-5836-0_5.

Erik Angner. Economists as experts: Overconfidence in theory and practice. *Journal of Economic Methodology*, 13(1):1–24, 2006. doi: 10.1080/13501780600566271. URL http://dx.doi.org/10.1080/13501780600566271.

Shahar Avin. Centralised funding and epistemic exploration. *The British Journal for the Philosophy of Science*, forthcoming. doi: 10.1093/bjps/axx059. URL http://dx.doi.org/10.1093/bjps/axx059.

Nachman Ben-Yehuda. Deviance in science: Towards the criminology of science. *British Journal of Criminology*, 26(1):1–27, 1986. doi: 10.1093/oxfordjournals.bjc.a047577. URL http://dx.doi.org/10.1093/oxfordjournals.bjc.a047577.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical

significance. *Nature Human Behaviour*, 2(1):6–10, 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0189-z. URL http://dx.doi.org/10.1038/s41562-017-0189-z.

Lutz Bornmann. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245, 2011. ISSN 1550-8382. doi: 10.1002/aris.2011.1440450112. URL http://dx.doi.org/10.1002/aris.2011.1440450112.

Thomas Boyer. Is a bird in the hand worth two in the bush? Or, whether scientists should publish intermediate results. *Synthese*, 191(1): 17–35, 2014. ISSN 0039-7857. doi: 10.1007/s11229-012-0242-4. URL http://dx.doi.org/10.1007/s11229-012-0242-4.

Thomas Boyer-Kassem and Cyrille Imbert. Scientific collaboration: Do two heads need to be more than twice better than one? *Philosophy of Science*, 82(4):667–688, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/682940.

John M. Braxton. Deviance from the norms of science: A test of control theory. *Research in Higher Education*, 31(5):461–476, 1990. doi: 10.1007/BF00992713. URL http://dx.doi.org/10.1007/BF00992713.

Liam Kofi Bright. On fraud. *Philosophical Studies*, 174(2):291–310, 2017a. ISSN 1573-0883. doi: 10.1007/s11098-016-0682-7. URL http://dx.doi.org/10.1007/s11098-016-0682-7.

Liam Kofi Bright. Decision theoretic model of the productivity gap. *Erkenntnis*, 82(2):421–442, 2017b. ISSN 1572-8420. doi: 10.1007/s10670-016-9826-6. URL http://dx.doi.org/10.1007/s10670-016-9826-6.

Liam Kofi Bright. Du Bois' democratic defence of the value free ideal. *Synthese*, 195(5):2227–2245, 2018. ISSN 1573-0964.

doi: 10.1007/s11229-017-1333-z. URL http://dx.doi.org/10.1007/s11229-017-1333-z.

Liam Kofi Bright, Haixin Dang, and Remco Heesen. A role for judgment aggregation in coauthoring scientific papers. *Erkenntnis*, 83(2):231–252, 2018. ISSN 1572-8420. doi: 10.1007/s10670-017-9887-1. URL http://dx.doi.org/10.1007/s10670-017-9887-1.

Justin Bruner and Cailin O'Connor. Power, bargaining, and collaboration. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 7, pages 135–157. Oxford University Press, Oxford, 2017.

Justin P. Bruner. Policing epistemic communities. *Episteme*, 10(4):403–416, Dec 2013. ISSN 1750-0117. doi: 10.1017/epi.2013.34. URL http://dx.doi.org/10.1017/epi.2013.34.

Erwin Chargaff. Triviality in science: A brief meditation on fashions. *Perspectives in Biology and Medicine*, 19(3):324–333, 1976. doi: 10.1353/pbm.1976.0011. URL http://dx.doi.org/10.1353/pbm.1976.0011.

Diana Crane. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 2(4):195–201, 1967. ISSN 00031232. URL http://www.jstor.org/stable/27701277.

Partha Dasgupta and Paul A. David. Toward a new economics of science. *Research Policy*, 23(5):487–521, 1994. ISSN 0048-7333. doi: 10.1016/0048-7333(94)01002-1. URL http://www.sciencedirect.com/science/article/pii/0048733394010021.

Margaret Eisenhart. The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, 32(2):241–255, 2002. ISSN 1573-1898. doi: 10.1023/A:1016082229411. URL http://dx.doi.org/10.1023/A:1016082229411.

Edzard Ernst, T. Saradeth, and Karl Ludwig Resch. Drawbacks of peer review. *Nature*, 363(6427):296, 1993. doi: 10.1038/363296a0. URL http://dx.doi.org/10.1038/363296a0.

Henry Etzkowitz, Stefan Fuchs, Namrata Gupta, Carol Kemelgor, and Marina Ranga. The coming gender revolution in science. In Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, editors, *The Handbook of Science and Technology Studies*, chapter 17, pages 403–428. MIT Press, Cambridge, third edition, 2008. ISBN 9780262083645.

Daniele Fanelli. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, 5(4):e10271, Apr 2010. doi: 10.1371/journal.pone.0010271. URL http://dx.doi.org/10.1371/journal.pone.0010271.

Jere R. Francis. The credibility and legitimation of science: A loss of faith in the scientific narrative. *Accountability in Research: Policies and Quality Assurance*, 1(1):5–22, 1989. doi: 10.1080/08989628908573770. URL http://dx.doi.org/10.1080/08989628908573770.

Alvin I. Goldman. *Knowledge in a Social World*. Oxford University Press, Oxford, 1999. ISBN 0198237774.

Timothy Gowers. The end of an error? *The Times Literary Supplement*, October 2017. URL https://www.the-tls.co.uk/articles/public/the-end-of-an-error-peer-review/. Editorial.

Paul M. Grant. Scientific credit and credibility. *Nature Materials*, 1:139–141, 2002. doi: 10.1038/nmat756. URL http://dx.doi.org/10.1038/nmat756.

Sandra Harding. "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3):331–349, 1995. doi: 10.1007/BF01064504. URL http://dx.doi.org/10.1007/BF01064504.

Remco Heesen. Academic superstars: Competent or lucky? *Synthese*, 194 (11):4499–4518, 2017a. ISSN 1573-0964. doi: 10.1007/s11229-016-1146-5. URL http://dx.doi.org/10.1007/s11229-016-1146-5.

Remco Heesen. Communism and the incentive to share in science. *Philosophy of Science*, 84(4):698–716, 2017b. ISSN 0031-8248. doi: 10.1086/693875. URL http://dx.doi.org/10.1086/693875.

Remco Heesen. Why the reward structure of science makes reproducibility problems inevitable. Manuscript, September 2017c. URL http://remcoheesen.files.wordpress.com/2015/03/rewards-and-reproducibility2.pdf.

Remco Heesen. When journal editors play favorites. *Philosophical Studies*, 175(4):831–858, 2018. ISSN 0031-8116. doi: 10.1007/s11098-017-0895-4. URL http://dx.doi.org/10.1007/s11098-017-0895-4.

Remco Heesen, Liam Kofi Bright, and Andrew Zucker. Vindicating methodological triangulation. *Synthese*, forthcoming. ISSN 1573-0964. doi: 10.1007/s11229-016-1294-7. URL http://dx.doi.org/10.1007/s11229-016-1294-7.

Erin Hengel. Publishing while female: Are women held to higher standards? Evidence from peer review. Manuscript, August 2018. URL http://www.erinhengel.com/research/publishing_female.pdf.

David L. Hull. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. The University of Chicago Press, Chicago, 1988. ISBN 0226360504.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, Aug 2005. doi: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

Saana Jukola. A social epistemological inquiry into biases in journal peer review. *Perspectives on Science*, 25(1):124–148, 2017. doi: 10.1162/POSC_a_00237. URL http://dx.doi.org/10.1162/POSC_a_00237.

J. Katzav and K. Vaesen. Pluralism and peer review in philosophy. *Philosophers' Imprint*, 17(19):1–20, 2017. URL http://hdl.handle.net/2027/spo.3521354.0017.019.

Philip Kitcher. The division of cognitive labor. *The Journal of Philosophy*, 87(1):5–22, 1990. ISSN 0022362X. URL http://www.jstor.org/stable/2026796.

Richard L. Kravitz, Peter Franks, Mitchell D. Feldman, Martha Gerrity, Cindy Byrne, and William M. Tierney. Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? *PLoS ONE*, 5(4):e10072, 2010. doi: 10.1371/journal.pone.0010072. URL http://dx.doi.org/10.1371/journal.pone.0010072.

Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, second edition, 1986.

Carole J. Lee. The limited effectiveness of prestige as an intervention on the health of medical journal publications. *Episteme*, 10(4):387–402, 2013. doi: 10.1017/epi.2013.35. URL http://dx.doi.org/10.1017/epi.2013.35.

Carole J. Lee. Revisiting current causes of women's underrepresentation in science. In Jennifer Saul and Michael Brownstein, editors, *Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology*, chapter 2.5, pages 265–282. Oxford University Press, Oxford, 2016. doi: 10.1093/acprof:oso/9780198713241.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780198713241.001.0001.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

Christin List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001. ISSN 1467-9760. doi: 10.1111/1467-9760.00128. URL http://dx.doi.org/10.1111/1467-9760.00128.

Helen E. Longino. *Science as Social Knowledge*. Princeton University Press, 1990.

Karen Seashore Louis, Lisa M. Jones, and Eric G. Campbell. Macroscope: Sharing in science. *American Scientist*, 90(4):304–307, 2002. ISSN 00030996. URL http://www.jstor.org/stable/27857685.

Bruce Macfarlane and Ming Cheng. Communism, universalism and disinterestedness: Re-examining contemporary support among academics for Merton's scientific norms. *Journal of Academic Ethics*, 6(1):67–78, 2008. ISSN 1570-1727. doi: 10.1007/s10805-008-9055-y. URL http://dx.doi.org/10.1007/s10805-008-9055-y.

Robert K. Merton. A note on science and democracy. *Journal of Legal and Political Sociology*, 1(1–2):115–126, 1942. Reprinted in Merton (1973, chapter 13).

Robert K. Merton. Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22(6):635–659, 1957. ISSN 00031224. URL http://www.jstor.org/stable/2089193. Reprinted in Merton (1973, chapter 14).

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63,

1968. ISSN 00368075. URL http://www.jstor.org/stable/1723414. Reprinted in Merton (1973, chapter 20).

Robert K. Merton. Behavior patterns of scientists. *The American Scholar*, 38 (2):197–225, 1969. ISSN 00030937. URL http://www.jstor.org/stable/41209646. Reprinted in Merton (1973, chapter 15).

Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago, 1973. ISBN 0226520919.

Brian A. Nosek, Jeffrey R. Spies, and Matt Motyl. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012. doi: 10.1177/1745691612459058. URL http://pps.sagepub.com/cgi/content/abstract/7/6/615.

Cailin O'Connor and Justin Bruner. Dynamics and diversity in epistemic communities. *Erkenntnis*, forthcoming. ISSN 1572-8420. doi: 10.1007/s10670-017-9950-y. URL http://dx.doi.org/10.1007/s10670-017-9950-y.

Slobodan Perović, Sandro Radovanović, Vlasta Sikimić, and Andrea Berber. Optimal research team composition: data envelopment analysis of Fermilab experiments. *Scientometrics*, 108(1):83–111, 2016. doi: 10.1007/s11192-016-1947-9. URL http://dx.doi.org/10.1007/s11192-016-1947-9.

Douglas P. Peters and Stephen J. Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2):187–195, 1982. doi: 10.1017/S0140525X00011213. URL http://dx.doi.org/10.1017/S0140525X00011213.

Katarina Prpić. Gender and productivity differentials in science. *Scientometrics*, 55(1):27–58, 2002. ISSN 0138-9130. doi: 10.1023/A:1016046819457. URL http://dx.doi.org/10.1023/A:1016046819457.

RIN. Activities, costs and funding flows in the scholarly communications system in the UK. Technical report, Cambridge Economic Policy Associates on behalf of the Research Information Network, 2008. URL http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/activities-costs-and-funding-flows-scholarly-commu.

Felipe Romero. Novelty versus replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, 84(5):1031–1043, 2017. ISSN 0031-8248. doi: 10.1086/694005. URL http://dx.doi.org/10.1086/694005.

Jennifer Saul. Implicit bias, stereotype threat, and women in philosophy. In Katrina Hutchison and Fiona Jenkins, editors, *Women in Philosophy: What Needs to Change?*, chapter 2, pages 39–60. Oxford University Press, Oxford, 2013.

Paul E. Smaldino and Richard McElreath. The natural selection of bad science. *Royal Society Open Science*, 3(9), 2016. doi: 10.1098/rsos.160384. URL http://rsos.royalsocietypublishing.org/content/3/9/160384.

Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182, 2006. URL http://jrs.sagepub.com/content/99/4/178.short.

Paula E. Stephan. The economics of science. *Journal of Economic Literature*, 34(3):1199–1235, 1996. URL http://www.jstor.org/stable/2729500.

Michael Strevens. The role of the priority rule in science. *The Journal of Philosophy*, 100(2):55–79, 2003. ISSN 0022362X. URL http://www.jstor.org/stable/3655792.

Michael Strevens. The role of the Matthew effect in science. *Studies in History and Philosophy of Science Part A*, 37(2):159–170, 2006. ISSN 0039-3681. doi: http://dx.doi.org/10.1016/j.shpsa.2005.07.009. URL http://www.sciencedirect.com/science/article/pii/S0039368106000252.

Michael Strevens. Herding and the quest for credit. *Journal of Economic Methodology*, 20(1):19–34, 2013. doi: 10.1080/1350178X.2013.774849. URL http://dx.doi.org/10.1080/1350178X.2013.774849.

Michael Strevens. Scientific sharing: Communism and the social contract. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 1. Oxford University Press, Oxford, 2017. URL https://philpapers.org/rec/STRSSC-2.

Johanna Thoma. The epistemic division of labor revisited. *Philosophy of Science*, 82(3):454–472, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/681768.

Virginia Valian. *Why So Slow? The Advancement of Women*. MIT Press, Cambridge, 1999. ISBN 9780262720311.

Richard Van Noorden. The true cost of science publishing. *Nature*, 495 (7442):426–429, 2013. ISSN 0028-0836. doi: 10.1038/495426a. URL http://dx.doi.org/10.1038/495426a.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/644786.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6(2):185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://dx.doi.org/10.3366/E174236000900063X.

Kevin J. S. Zollman. The credit economy and the economic rationality of science. *The Journal of Philosophy*, 115(1):5–33, 2018. doi: 10.5840/jphil201811511. URL http://dx.doi.org/10.5840/jphil201811511.

Harriet Zuckerman and Jonathan R. Cole. Women in American science. *Minerva*, 13(1):82–102, 1975. ISSN 1573-1871. doi: 10.1007/BF01096243. URL http://dx.doi.org/10.1007/BF01096243.

Epistemic Loops and Measurement Realism

Alistair M. C. Isaac

Abstract

Recent philosophy of measurement has emphasized the existence of both diachronic and synchronic "loops," or feedback processes, in the epistemic achievements of measurement. A widespread response has been to conclude that measurement outcomes do not convey interest-independent facts about the world, and that only a coherentist epistemology of measurement is viable. In contrast, I argue that a form of measurement realism is consistent with these results. The insight is that antecedent structure in measuring spaces constrains our empirical procedures such that successful measurement conveys a limited, but veridical knowledge of "fixed points," or stable, interest-independent features of the world.

§1 Introduction

Recent philosophy of measurement has employed detailed case studies to highlight the complex, iterative process by which measurement practices are refined. Typically, these examples are taken to support some form of epistemic coherentism, on which the validation of measurement procedures, and thus their epistemic import, is irreducibly infected by the contingent history of their development in aid of human interests. This coherentism in turn undermines *measurement realism*, the view that outcomes of successful measurement practices veridically represent objective (i.e. interest-independent) features of the world. For instance, van Fraassen (2008) takes the historical contingency of measurement practice to support empiricism, and Chang (2012) argues that only a pragmatic, interest-relative "realism" about measurement outcomes is plausible, not one which interprets them as corresponding to objective features in the world. More generally, Tal (2013) identifies coherentism as a major trend within contemporary philosophy of measurement.

I argue that the iterative and coherentist features of measurement practice these authors rightly emphasize are nevertheless consistent with realism about measurement outcomes. Nevertheless, my position contrasts significantly with that of other measurement realists, such as Byerly and Lazara (1973) or Michell (2005), who take measurement realism to be continuous with global scientific realism. On their view, measurement realism is a *stronger* position than traditional realism, imputing reality not only to theoretical objects and laws, but also to their quantitative character. The view defended here reverses this priority, articulating a realism about measurement outcomes *weaker* than traditional realism. In particular, I argue that the convergent assignment of increasingly precise values that constitutes successful measurement serves as incontrovertible evidence for *fixed points* in the world — features or events standing in stable quantitative relationships — even though the evidence it provides for any non-numerical theoretical description of these points is defeasible. The insight here is that measurement is more evidentially demanding than traditional confirmation, i.e. it requires a greater contribution from the interest-independent world to succeed than mere qualitative experiments. I argue that this greater evidential demand is a consequence of the

antecedent numerical structure in which measurement outcomes are represented. This antecedent structure blocks the possibility of gerrymandered categories that crosscut the joints of nature. Consequently, successful measurement constitutes a substantive enough epistemic achievement that we may legitimately "factor out" the contribution to success made by human interests, and accept the outcome as representing an objective feature of the world.

After surveying the motivations for measurement coherentism, I elaborate on the notion of "successful" measurement, and why it poses a challenge to coherentism. The paper concludes with a more careful articulation of the distinctive features of fixed point realism.

§2 Epistemic Loops in Measurement Practice

Contemporary measurement coherentism is motivated by two types of case study, each identifying a different kind of epistemic "loop," or feedback process driving knowledge formation. Chang and van Fraassen emphasize diachronic examples of epistemic iteration, where the feedback process extends over several stages of mutual influence between theory change and refinement of measurement practice. A different kind of epistemic loop has been discussed by Tal and metrologist Mari, who highlight the role of models in the calibration of measurement instruments and the assignment of quantity values, illustrating a synchronic epistemic interdependence between theory and measurement.

§2.1 Epistemic Iteration

Chang (2004) defines *epistemic iteration* as "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals" (45). He takes this process to support a "progressive coherentism": on the one hand, the criteria for measurement success are internal to a practice, so scientific knowledge does not rest on an independent foundation; on the other hand, these internal criteria may be used to evaluate new practices as improvements or refinements on their predecessors, thereby allowing for scientific progress (in contrast to traditional coherentism, Chang 2007). In the context of measurement, this means that later measurement practices may be understood as in some sense "better" than earlier ones, yet these "epistemic achievements" should not be cashed out as greater degree of correspondence to quantities in the world.

For instance, thermometry as a practice begins with subjective assignments of relative heat on the basis of our bodily experiences. Noticing that fluids appear to change volume in rough correspondence with these subjective sensations, one may construct a thermoscope, or device allowing comparison of relative fluid volumes in different circumstances. Already a theoretical leap is required to identify the cause of these changes in relative volume with the cause of our differing subjective sensations, especially given the discrepancies between these sensations and our thermoscopic readings (e.g. contrary to experience, caves are warmer in summer than they are in winter). Nevertheless, the move to the thermoscope constitutes an epistemic achievement, in the sense that it allows for greater regularity in the assignment of relative temperatures, both

across contexts and across observers. A similar pattern is seen in the move from thermoscope to thermometer, which enables assignment of numbers to temperatures. Numerical representation constitutes a yet greater epistemic achievement, insofar as it allows comparison of temperature assignments across devices. Nevertheless, this practice does not itself guarantee greater veracity of temperature assignments, since it rests on the assumption that temperature varies linearly with changes in the height of thermometric fluid. But this assumption cannot itself be verified, as that would require access to temperature in the world by some means independent of thermometry. Similar achievements, (seemingly) inextricably entangled with theory, may be seen at each further stage in the development of thermometric practice.

The moral of this case study is the historical contingency of thermometry, and thus of its results. At each stage in the development of thermometry, an advance in theory was required to extend measurement practice. Internal criteria of consistency and increased precision in the assignment of numerical values establish the new practice as an advance over the previous one. Yet, the application of these criteria is not empirically constrained. When one assumes that "temperature" (whatever it may be) varies linearly with changes in the height of the indicator column in an air thermometer, one is making an assumption both necessary for measurement progress and in principle non-empirical, since no independent access to "temperature," outside the behavior of the very devices and procedures under investigation, is possible: "*Prior to the construction of a thermometer, there is no thermometer to settle that question!*" (van Fraassen 2008, 126, emphasis in original). Chang (2004) argues that, in order to make sense of the "progress" exemplified by cases like these, we have to "look away from truth," and appeal only to historically contingent criteria for success (227)—"scientific progress … cannot mean closer approach to the truth" (228); "Truth, in the sense of correspondence to reality, is beyond our reach" (Chang 2007, 20). The delusion that one may evaluate the correspondence between our assignment of temperatures and the objective state of the world rests on the mistaken and "impossible god-like view in which nature and theory and measurement practice are all accessed independently of each other" (van Fraassen 2008, 139). Rather, the only relevant notion of "truth" for assessing the success of thermometry "rests first and foremost on coherence with the rest of the system" (Chang 2012, 242).

§2.2 Models and Calibration

Another kind of epistemic loop is found in synchronic measurement practice, where *models* play a constitutive role in determining measurement outcomes. The crucial concept here is *calibration*, the process of correcting a measurement device for inferred discrepancies between its readout and the target value. Calibration is a necessary feature of all sophisticated measurement, yet the process of calibration illustrates the ineliminable role of theoretical posits in the very assignment of quantity values in an act of measurement. When measuring, scientists do not (as one might naively suppose) read values directly from nature, rather they employ models of the interaction between measurement device and target system in order to "correct" the readout value to a final assigned value (Mari and Giordani 2014).

Tal (2014) illustrates this point through the example of the measurement of time, in particular coordinated universal time (UTC). The second is presently defined as 9,192,631,770 periods of the hyperfine transition between the two ground states of a caesium-133 atom at zero degrees Kelvin.[1] Models feature at every step of the process leading from devices that interact directly with caesium atoms to the UTC. First, it is impossible to probe caesium atoms at absolute zero, so the enumeration of hyperfine transitions output by a caesium clock must be corrected for this discrepancy. This, as well as other corrections, rely on models of the physical interaction between the device and the atom in order to infer the discrepancy between the actual state of the system and the idealized state referred to in the definition. Caesium clocks are too complex to run continuously, so their output is used to calibrate more mundane atomic clocks (301). Furthermore, the UTC itself is not identified with the output of any one clock; rather, it is calculated retrospectively by a weighted average over all participating atomic clocks, with weights determined by the degree of past fit between each clock and previous calculations of UTC (302–3).

The lessons of this example are analogous to those of epistemic iteration: measurement improvement appears to rest on internal standards of coherence rather than on correspondence with external quantities. The weighting procedure that leads to UTC, for instance, "promotes clocks that are stable relative to each other" (304). Success at achieving this stability indeed demonstrates "genuine empirical knowledge," but not knowledge in the first instance about a regularity in the objective world, but rather a regularity "in the behaviour of instruments" (327). Consequently, it is a "conceptual mistake" to think that "the stability of measurement standards can be analysed into distinct contributions by humans and nature" (328). On an extreme interpretation of this view, even computer simulation constitutes a form of measurement (Morrison 2009). The basic idea is that, once we grant the ineliminable role of models in measurement, it is a small conceptual step to accept that the aspect of measurement involving empirical contact with the world may be arbitrarily distant from that involving modeling (Parker 2017).

§3 Achieving Successful Measurement

For the remainder of this paper, I wish to grant the basic descriptive features of this account: both diachronically and synchronically, successful measurement involves epistemic loops. Nevertheless, I will argue, there is a form of measurement realism consistent with these loops; one on which the contingent, interest-relative, and theory-laden aspects of measurement may indeed be factored out, leaving the bare, objective facts about the world conveyed by successful measurement.

[1] Arguably, the process of establishing UTC is not measurement at all — since the length of the second is *defined* by caesium-133 transitions, it is not subject to empirical determination. The purpose of the project Tal examines is not to establish a value, as in paradigmatic cases of measurement, but rather to coordinate time-relevant activities across the globe with maximal precision. I set this concern aside for the discussion here, since Tal's analysis has been so influential in philosophy of measurement, and his conclusions concerning the model-mediation of measurement incontrovertibly reflect the practices of metrologists.

Seattle, WA; 1-4 November 2018 -345-

But what is "successful measurement"? For the purposes of discussion here, I take *measurement* to be any empirical procedure for assigning points (or regions) in a metric space to states of the world, where a *metric space* is any set of elements with a distance metric defined over it. This means, on the one hand, that I rule out degenerative forms of "measurement" that simply assign objects to categories, or place them in an order (the *nominal* and *ordinal* scales of Stevens 1946). On the other hand, I include measurement procedures that map states of the world into any geometrical space, not just the real line, so long as they have an assigned distance metric (siding with Suppes, et al. 1989, against Díez 1997); nevertheless, in the interests of simplicity, I will refer to these outcomes as "numerical" assignments, since they may be represented by vectors of real numbers. In line with Krantz et al. (1971), I take it that one can determine whether or not an empirical procedure constitutively requires the metric features of a geometrical space by analyzing whether these remain invariant across permissible transformations over the mapping into that space.[2]

I take *successful* measurement to exhibit two key features: *convergence* and *precision*. These features pose a significant challenge to the thoroughgoing coherentist.

§3.1 Convergence

Coherentists have emphasized the theory-ladenness of both diachronic and synchronic aspects of measurement refinement. However, a hallmark of sophisticated scientific measurement is its attempt to factor out the role of theory in measurement by employing different theoretical commitments to measure the same quantity. A measurement practice *converges* when procedures employing different theoretical commitments arrive at the same outcome.

For instance, in the early 20th century, a wide variety of phenomena were investigated, employing distinct methods and theoretical commitments, in the attempt to measure Avogadro's constant N_A, the number of particles in a mole of substance. Perhaps most well-known are Perrin's experiments on Brownian motion, which, in combination with Einstein's theoretical analysis, allowed an assignment of value to N_A. However, similar values were achieved by radically different means. For instance, Millikan was able to determine N_A by measuring charge of the electron through his oil drop experiments and dividing the Faraday constant (charge of a mole of electrons) by his result. Millikan's measurement relied on Stokes' theoretical analysis of the movement of spheres through a viscous fluid — insofar as Brownian motion was a factor, it was as a source of noise, not (as for Perrin) a source of evidence. Black body radiation and the blue

[2] For instance, consider two procedures for assigning real numbers to my students. On the first, I assign a number to each letter-type with which a student's name begins (e.g. A=1, B=3,…); on the second, I hold a meter stick up to each student and note their height. The former procedure is indifferent to the algebraic structure of the real line (letters do not add or subtract from each other systematically), and thus metric features of the real line are not invariant across alternative, equally permissible assignments of numbers (e.g. A=7, B=15,…). The second does make use of algebraic structure (as heights do "add" through concatenation), and thus metric features remain invariant across alternative assignments (Jamal is twice the height of Leslie, whether their heights are represented in inches or centimeters). So, on the present definition, the latter procedure is measurement, but the former is not.

of the sky are examples of other phenomena that, when combined with theoretical models of photon emission and diffraction respectively, allow alternate means of measuring N_A. Insofar as these procedures assign the same value to N_A, they converge.

I want to stress that the point being made here is *not* the traditional realist one, that these practices provide converging evidence for the particulate nature of matter, whether as "common cause" (Salmon 1984) or most likely hypothesis (Psillos 2011). Those arguments are instances of *abduction*, while I am interested in whether a stronger, non-abductive conclusion may be drawn from convergence. A better analogy is with the discussion of robustness in the modeling literature: a result is *robust* if it is obtained by a plurality of models that each make different simplifying assumptions (Weisberg 2006). The particulate nature of matter is not robust in this sense across different measurement practices, since it is assumed by all of them. However, the value of N_A is robust, since that value is not itself assumed, and is obtained with a great degree of agreement despite differences in the assumptions made by each measurement practice (and its supporting models). I claim that convergence toward this value provides robust, non-abductive evidence for an objective feature of the world.

This example is in no way exceptional: convergent measurement practices are rife across the sciences. Smith and Miyake, for instance, have investigated a number of examples. Thomson's convergent measurements of the charge of the electron employed a variety of different methods and assumptions (Smith 2001). Early attempts to measure the density of the interior of the earth likewise assumed a variety of different theoretical models (Miyake 2018). In more recent research, measurements of the constants that govern molecular vibration converge across spectroscopy, chemistry, thermodynamics, and femtochemistry (Smith and Miyake, *manuscript*). To pick an example from an entirely different area of science, measurements of the spectral sensitivity of mammalian retinal receptors employing psychophysical methods (extracting sensitivity curves from behavioral color matching experiments, as performed by Helmholtz in the late 19th century) converge closely with 20th century physiological methods (detecting rate of nerve firing in (e.g.) cow retinal tissue in response to single wavelength lights, Wandell 1995). In all of these cases, "What is being shown through the convergence of these measurements is that the discrepancies between the different measurements … are due to the particularities of the models being used" (Miyake, 2018, 336). In other words, convergence factors out model-sensitive features of measurement; in order for it to occur, "the empirical world has to cooperate" (Smith 2001, 26).

§3.2 Precision

Traditionally, measurement success was evaluated with respect to two features: accuracy and precision. *Accuracy* was degree of approach to true value, while *precision* was degree of specificity in the value provided. The considerations in §2 undermine the criterion of accuracy, since they show we have no independent access to "true values" and thus cannot use them as standards for evaluating measurement (Mari 2003). Nevertheless, we can still assess measurements for precision, since it may be defined operationally: a measurement is *precise* to the

extent that it returns the same result when performed repeatedly. The number of *significant figures* in a numerical assignment indicates the degree of measurement precision, since these characterize the size of the region within which repeated measurements fall.

Coherentists stress the fact that increased precision is a purely internal criterion for improving measurement. Here, however, I want to stress the way in which increased precision constitutes a qualitatively different, and more impressive, epistemic achievement than other forms of empirical success, such as qualitative prediction or improved coherence of classification. These qualitative achievements are subject to worries about semantic and theoretical holism: one may always succeed in classification, or correct qualitative prediction, by suitably redrawing the boundaries of one's theoretical concepts. As LaPorte (2004) argues, when faced with anomalies in the relationship between guinea pigs and prototypical rodents, or birds and dinosaurs, scientists face a *choice* whether to expand or contract their previous categories to include or exclude perceived outliers (a similar case is made by Slater 2017 for Pluto and planethood). Nothing about the prior conceptual framework itself forces this choice one way or another, nor do demands for internal consistency.

Measurement is different from mere categorization precisely because it maps states into a metric space. The crucial point to note here is that a metric space has *antecedent structure*: the distances between points on the real line, and the algebraic relationships between them, are fixed *before* we employ it to represent height or temperature or electric charge. This antecedent structure constrains the relationship between measurement outcomes, independently restricting our assessment of them as same or different, or converging or not, in a manner impervious to ad hoc revision. Increase in precision occurs when successive measurement practices are able to shrink distances (between repeated measurements within each practice) determined by the metric of the representing space. Thus, the metric of this space serves two functions: (i) it represents the distances between different measured quantities, but (ii) it also provides a directed metric for improving measurement of a single quantity, since it determines the distances between repeated measurements that characterizes their precision. Consequently, pace van Fraassen, attempts to increase precision are empirically constrained, since this directed metric for improvement can only be satisfied through the cooperation of nature: if nature is not sufficiently stable where we probe it, no choice, convention, or increased coherence can reduce the distances between our repeated attempts to measure it. Some examples will illustrate this point.

Consider, for instance, determinations of the boiling point of water. Chang (2004, Ch. 1) surveys the sequence of choice points in the early practice of thermometry leading to relative stability in the measurement of this temperature: what are the visual indicators of boiling, where should the thermometer be positioned, what should be the shape of the vessel holding the water, its material, etc.[3] Decisions on each of these points affect the relative stability in the thermometric reading, illustrating the naivety of a view on which

[3] The issue here is the phenomenon of "superheating," whereby water with relatively little dissolved gas, or in a flask with very small surface area, may be heated to a higher temperature without bubbling.

boiling point is a simple phenomena merely waiting to be observed. Nevertheless, in committing to represent the boiling point numerically, investigators subjected themselves to a criterion for success distinct from coherence. If the numbers assigned by thermometers within this-shaped vessels and that-shaped ones differ during phenomenologically similar bubblings, then the distance between those numbers provides a criterion of difference that must be respected if thermometric practice is to count as measurement. Restricting attention to those vessels that minimize distances between numerical outcomes is thus not a mere choice, or gerrymandering of the category "boiling," since it is forced upon the investigator by an antecedent metric for success.

Likewise, consider again the determination of UTC through the retrospective weighting of the comparison set of atomic clocks. For Tal, the success of this procedure is evidence for stability in our clocks, but not for any human-independent feature of the world. Nevertheless, UTC is constrained by the world in two distinct ways. First, through empirical contact with caesium atoms. While this contact is mediated by models, these models themselves are the result of convergent measurements of atomic phenomena through a wide variety of means, employing distinct theoretical assumptions. Second, the distance metric of the real line constrains the assessment of fit between clocks in the set. While the algorithm that weights them takes degree of internal agreement as the standard for higher weighting, the metrical structure of the space in which relative rates of the clocks are assessed ensures relative agreement cannot be stipulated, fudged, or gerrymandered. The clocks need to cooperate by performing stably enough that they may be compared with a high degree of precision, and this stable point remains tethered to a robust regularity in the world through checks with the convergent behavior of caesium.

While UTC is in some respects atypical (see footnote 1), these three features — internal coordination of outcomes, empirical checks, and directed improvement constrained by the real line — are features of scientific measurement in general. What Tal's discussion of the UTC obscures is the sheer number of empirical checks typically involved, and the strictness of the demands placed by conformity to the metric of improvement the measuring space provides. In official determinations of fundamental physical constants, convergence is demanded across *all* measurement procedures, as assessed by the law-governed interrelationship between physical quantities, and the degree of precision achieved illustrates the strictness of this demand. For instance, in late 19th century measurements of N_A by Perrin and e (charge of electron) by Thomson, only 2 to 3 significant figures were typically obtained within method, and convergence across methods often only agreed as to order of magnitude. By 1911, Millikan was measuring both e and N_A to 4 significant figures, and demonstrating that the models employed to calibrate the oil drop method converged closely with other aspects of physical theory (1911). As of 2014, N_A was being measured at upwards of 9 significant figures, and e upwards of 11 (Mohr et al. 2016).[4] In each case, the increase in precision has been constrained by the antecedent structure of the real line, and thus is not itself a matter of mere convention or coherence. Rather, the world must cooperate by remaining

[4] It is expected that after the 2018 26th General Conference on Weights and Measures, N_A and e will be fixed as constants to which other quantities may be referred during measurement.

Seattle, WA; 1-4 November 2018 -349-

sufficiently stable if such precision is to be possible; consequently, precise values constitute robust evidence for points of objective fixity in the world revealed through measurement.

§4 Conclusion: Fixed-Point Realism

Traditional scientific realism rests on an abductive inference from observed empirical success to presumed underlying causes. Successful measurement may certainly be used in such an inference, but I claim here that it non-abductively supports a more modest realism:

> **Fixed Point Realism** – values obtained through successful measurement veridically represent objective fixed points in the world, which may be exhaustively characterized by the pattern of distances that obtain between them in a metric space.

FPR is a form of *epistemic structural realism*. It differs from traditional realism insofar as it claims a veridical characterization of the world is possible independent of any particular theoretical description. Our theory of the nature of temperature or of state changes may change radically, yet the points of relative stability characterizing, e.g., boiling point of water, "absolute zero," freezing point of oxygen, etc., will stay robust across any such change, and that robustness may be represented by their relative positions within a numerical scale.

FPR differs from other flavors of structural realism in the type of structure to which it is committed. Structural realists typically focus on the rich mathematical structure of physical theory, and derivation or limit relations that hold between successive theories, e.g. Newton's laws are a limit case of relativistic mechanics (Worrall 1989). FPR commits itself only to *geometric* structure, i.e. the pattern of relative distances that obtain between points of stability as represented in a metric space. Just as our theoretical description of these stable points may change, so may our mathematical account of their relationship — if new mathematical physics fails to derive old equations as limit cases, this in no way jeopardizes the veridicality of this geometric structure.

Finally, FPR disagrees with coherentism, insofar as it asserts that the geometrical structure uncovered through acts of successive measurement obtains in the world independent of our practices. It does not deny the importance of epistemic loops for understanding the process of measurement. Nevertheless, it takes convergence in measured values to indicate that the points of stability they represent obtain independent of the theoretical commitments encapsulated in the models used for calibration. Likewise, it takes increased precision to constitute a criterion for measurement success over and above that of coherence, one that is only realized when the interest-independent world cooperates with us by remaining stable when we probe it.

Bibliography

Byerly, H., and V. Lazara (1973) "Realist Foundations of Measurement," *Philosophy of Science* 40:10–28.

Chang, H. (2004) *Inventing Temperature*, Oxford UP.

Chang, H. (2007) "Scientific Progress: Beyond Foundationalism and Coherentism," O'Hear (ed.) *Royal Institute of Philosophy Supplement* 61:1–20.

Chang, H. (2012) *Is Water H_2O?* Springer.

Díez, J. (1997) "A Hundred Years of Numbers: An Historical Introduction to Measurement Theory 1887–1990, part ii," *Studies in History and Philosophy of Science* 28:237–265.

Krantz, D., R. Luce, P. Suppes, and A. Tversky (1971) *Foundations of Measurement*, vol. 1, Dover.

LaPorte, J. (2004) *Natural Kinds and Conceptual Change*, Cambridge UP.

Mari, L. (2003) "Epistemology of Measurement," *Measurement* 34:17–30.

Mari, L., and A. Giordani (2014) "Modeling Measurement: Error and Uncertainty," in Boumans, Hon, and Petersen (eds.) *Error and Uncertainty in Scientific Practice*, Pickering & Chatto: 79–96.

Michell, J. (2005) "The Logic of Measurement: A Realist Overview," *Measurement* 38:285–294.

Millikan, R. (1911) "On the Elementary Electrical Charge and the Avogadro Constant," *Physical Review* 2:349–397.

Miyake, T. (2018) "Scientific Realism and the Earth Sciences," in Saatsi (ed.) *The Routledge Handbook of Scientific Realism*, Routledge: 333–344.

Mohr, P., D. Newell, and B. Taylor (2016) "CODATA Recommended Values of the Fundamental Physical Constants: 2014," *Review of Modern Physics* 88:035009.

Morrison, M. (2009) "Models, Measurement and Computer Simulation: The Changing Face of Experimentation," *Philosophical Studies* 143:33–57.

Parker, W. (2017) "Computer Simulation, Measurement, and Data Assimilation," *British Journal for Philosophy of Science* 68:273–304.

Psillos, S. (2011) "Moving Molecules above the Scientific Horizon: On Perrin's Case for Realism," *Journal for General Philosophy of Science* 42:339–363.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton UP.

Slater, M. (2017) "Plato and the Platypus: An Odd Ball and an Odd Duck – On Classificatory Norms," *Studies in History and Philosophy of Science* 61:1–10.

Smith, G. (2001) "J.J. Thomson and the Electron, 1897–1899," in Buchwald and Warwick (eds.) *Histories of the Electron*, MIT Press.

Smith, G., and T. Miyake (*manuscript*) "Realism, Physical Meaningfulness, and Molecular Spectroscopy"

Stevens, S. (1946) "On the Theory of Scales of Measurement," *Science* 103(2684):677–680.

Suppes, P., D. Krantz, R. Luce, and A. Tversky (1989) *Foundations of Measurement*, vol. 2, Dover.

Tal, E. (2013) "Old and New Problems in Philosophy of Measurement," *Philosophy Compass* 8/12:1159–1173.

Tal, E. (2014) "Making Time: A Study in the Epistemology of Measurement," *British Journal for Philosophy of Science* 67:297–335.

Wandell, B. (1995) *Foundations of Vision*, Sinauer.

Weisberg, M. (2006) "Robustness Analysis," *Philosophy of Science* 73:730–742.

Worrall, J. (1989) "Structural Realism: The Best of Both Worlds," *Dialectica* 43:99–124.

van Fraassen, B. (2008) *Scientific Representation*, Oxford UP.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -352-

The relationship between intervention and representation is currently resurfacing in philosophy of science. Analytical treatments of the specific intersections between *representation* and *intervention* have recently been explored in Hacking (1983), Radder (2003), Heidelberger (2003), van Fraassen (2008), and Keyser (2017). These accounts analyze intervention-based experimental and measurement practice and the *consequences* for representing and model-building. Of particular interest in my discussion is that some of these accounts explicitly differentiate between representational and productive roles in scientific practice. For example, Heidelberger (2003) and van Fraassen (2008) discuss the representational and productive roles of instruments in experiment and measurement. In the former role, relations in a natural phenomenon are represented in an instrument (van Fraassen 2008, 94). In the latter role, instruments create new phenomena or mimetic phenomena, which resemble natural phenomena. Keyser (2017) takes the distinction between representation and production a step further to differentiate two types of experimental/measurement methodologies:

> When scientists measure/experiment they can *take* measurements, in which case the primary aim is to represent natural phenomena. Scientists can also *make* measurements, in which case the aim is to intervene in order to *produce* experimental objects and processes—characterized as 'effects'. (Keyser 2017, 2)

On Keyser's account 'taking a measurement' involves a scientist using a result in the context of theory to represent a given phenomenon (2017, 9-15). In contrast, 'making a measurement' involves setting up experimental conditions to produce a phenomenon—where that phenomenon can be realized in nature but it can also be a brand new

phenomenon (Keyser 2017, 10). The difference between these two methodologies seems to be a matter of passive representation of a phenomenon vs. active intervention to produce a phenomenon. While the distinction between representation and intervention has been useful in classifying methodology in well-documented contexts like thermometry, microscopy, and cellular measurement, I argue that it falls apart in contexts where taking and making are *entangled*—such as in the context of biomarker measurement in the biomedical sciences.

In this discussion, I aim to show that in *complex methodological contexts*, representational and intervention-based roles require re-conceptualization. I analyze the *relations* between representation and intervention by focusing on the role of intervention in *mediating* representations. In Section 2, I show how applied scientific practice challenges the simple distinction between representational and intervention-based roles of experiment/measurement. In Section 3, I discuss the complex interaction between representation and intervention applied to methodology in biomarker measurement.

2. Methodology at the Intersection between Intervention and Representation

In order to understand why the distinction between representation and intervention needs a multifaceted approach, it is important to be explicit about what it means to represent and intervene in scientific practice. In Section 2.1, I draw on van Fraassen (2008) to discuss representation and both van Fraassen (2008) and Keyser (2017) to discuss intervention. Then in Section 2.2, I show how applied scientific practice challenges the simplistic distinction between representational and intervention-based

roles of experiment/measurement. I argue that the distinction between intervention and representation is less about *specific types of methodologies* in measurement/experiment and more about where one philosophically partitions the measurement *process*.

2.1. Representation and intervention

In experimental and measurement practice, representation has at least three important components: First, instruments or experimental contexts yield measurement values; Second, those values can only be interpreted within the context of a well-developed theory; and third, the relation between the measurement values and the phenomenon is determined by a user (e.g., experimenter). Van Fraassen (2008) provides a rich characterization of representation in measurement and experiment, which requires careful analysis. Worth noting is that van Fraassen takes measurements to be a "special elements of the experimental procedure" (2008, 93-94). For my discussion the embeddedness of measurement in experiment is not important. I will focus on the roles or processes within measurement and experimental practice. But to do this, I will sometimes refer to 'measurement' and other times to 'experiment'. Van Fraassen's characterization focuses on interaction and representation in measurement:

> A measurement is a physical interaction, set up by agents, in a way that allows them to gather information. The outcome of a measurement provides a representation of the entity (object, event, process) measured, selectively, by displaying values of some physical parameters that—according to the theory governing this context—characterize that object. (2008, 179-180)

Seattle, WA; 1-4 November 2018 -355-

For van Fraassen, measurement interaction between an object of measurement and apparatus generates a physical outcome—the "measurement outcome" or "physical correlate of the measurement outcome"—, which provides information content about the target of measurement (2008, 143). The contents of measurement outcomes convey information about *what is measured* through the mediation of theory. Van Fraassen posits that theoretical characterization of measurement interaction requires 'coherence':

> The theoretical characterization of the measurement situations is required to be coherent with the claims about the existence of measurement outcomes, their relation to what is measured, and their function as sources of information. (2008, 145)

In short, the theory tells a coherence story about "how its outcomes provide information about what is being measured" (145). Furthermore, the information content is representational. Van Fraassen says, "The outcome provides a representation *of* the measured item, but also represents it *as* thus or so" (2008, 180). To understand how the representational relation works, it is important to refer to van Fraassen's 'representation criterion':

> The criterion for what sorts of interactions can be measurements will be, roughly speaking, that the outcome must represent the target in a certain fashion—, selectively resembling it at a certain level of abstraction, according to the theory—*it is a representation criterion*. (van Fraassen 2008, 141).

Two aspects of the representation criterion require explanation: First, the distinction between "target" and "outcome"; and second, the role of theory in the operation of measurement. I begin with the former. Van Fraassen makes a technical

distinction between the target of measurement ('phenomena') and the outcome of measurement ('appearances'):

> Phenomena are observable, but their appearance, that is to say, *what they look like in given measurement or observation set-ups*, is to be distinguished from them as much as any person's appearance is to be distinguished from that person. (2008, 285)

For van Fraassen, phenomena are observable objects, events, and processes (2008, 283). He emphasizes that phenomena include all observable entities—whether observed or not (2008, 307). A given phenomenon can be measured in many different ways. The outcome of each measurement provides a perspective on a given phenomenon—meaning that the content of measurement tells us what things *look like*, not what they *are like* (2008, 176, 182). The *content* of the measurement outcome is an appearance.

An important qualification is that for van Fraassen, a representation does not represent on its own. The scientist selects the aspects/respects and degrees to which a representation represents a target. This relation can be expressed as: Z uses X to represent Y as F, for purposes P.

Now that the target and outcome of measurement have been characterized, we can specify van Fraassen's role of theory in measurement. According to van Fraassen, "Measurement is an operation that locates an item (already classified as in the domain of a given theory) in a logical space, provided by the theory to represent a range of possible states or characteristics of such items (164). Three things are worth noting about van Fraassen's discussion of logical spaces. First, a logical space provides a multidimensional mathematical space that locates potential objects of measurement (2008, 164). By

Seattle, WA; 1-4 November 2018 -357-

measuring we assign the item a location in a logical space. However, according to van Fraassen, it does not have to be on a real number continuum. As van Fraassen points out, items may be classified (by theory) on a range that is "an algebra", "lattice", or a "rudimentary poset" (2008, 172). Second, theoretical location depends on a "family of models" and not just an individual model (2008, 164). Third, an item is located in a "region" of logical space rather than at an exact point (2008, 165). Simply put, theory provides a classificatory system for what is measured. Importantly, theory is *necessary* for this type of classification. Van Fraassen says, "A claim of the form "This is an X-measurement of quantity M pertaining to S" makes sense *only* in a context where the object measured is already classified as a system characterized by quantity M" (2008, 144 my emphasis).

We can summarize the above discussion into four conditions for van Fraassen's account of representation in measurement/experiment practice:

i. Physical Interaction Condition: The interaction between apparatus an object produces a physical correlate of the measurement outcome.

ii. Theoretical Characterization Condition: The content of the measurement outcome is given a location in a logical space, which is governed by a family of theoretical models. An item's location within a logical space can change in content and truth conditions as accepted theories change.

iii. Representational Content Condition: The content of a measurement outcome provides a selective representation of a given target of measurement (phenomenon). Because representations do not represent on their own, users and pragmatic considerations set the representational relation such that: Z uses X to represent Y as F, for purposes P.

iv. Perspectival Information Condition: Measurement generates appearances, which are public, intersubjective, contents of measurement outcomes. Appearances provide selective information about phenomena. Thus information from measurement tells us what something *looks* like and not what something *is* like.

Van Fraassen notes that measurement and experiment are not only limited to a representational role, they can take on at least two productive roles. First, instruments can produce phenomena that "imitate" natural phenomena. That is, carefully controlled conditions give rise to mimetic effects that are used by scientists in the context of theory to resemble natural phenomena (2008, 94-95). It is important to note that van Fraassen emphasizes that natural phenomena are phenomena that exist *independent of human intervention* (2008, 95). The second productive role of instruments is that they are used as "engines of creation" to produce or manufacture new phenomena. Van Fraassen is not explicit about whether or not the representational roles can smear with the productive roles. There is no reason to assume that these roles cannot be combined; but that requires explicit philosophical work to see *how*, which I develop in Section 3.

Seattle, WA; 1-4 November 2018 -359-

Keyser (2017) is explicit about the relationship between the representational and intervention-based roles in science. He discusses the *use* of intervention for developing causal representations. Scientists intervene, thereby manipulating causal conditions within a given measurement or experimental system, which he calls 'intervention systems', to produce some sort of "effect" (Keyser 2017, 9-10). According to Keyser, "Intervention systems consist of organized experimental conditions and as such the effects that emerge are often sensitive to changes in conditions*"* (Keyser 2017, 10). Once a given effect is produced it can be used in order to be informative about causal relations for theoretical model building.

Keyser (2017) also differentiates between the methodologies of taking measurements vs. making measurements. I interpret that taking measurements involves three components: First, some instrument or experimental arrangement yields a qualitative or quantitative value; second, a 'theoretical representational framework'—which is just a body of models—is necessary in order to characterize that value according to parameters and relations between parameters; and third, a scientist sets up the resemblance relation between the measurement/experiment value and some aspect(s) of a phenomenon (Keyser 2017, 14-15). In contrast, when scientists make measurements they manipulate causal conditions—such as, preparatory, instrument, and background conditions—within an intervention system. This manipulation gives rise to some effect (Keyser 2017, 3-12).

There is something puzzling about Keyser's distinction between making vs. taking, if we apply the aforementioned conditions (i-iv): i. *Physical Interaction Condition*; ii. *Theoretical Characterization Condition*; iii. *Representational Content*

Condition; and iv. *Perspectival Information Condition*. Namely, it seems that 'making measurements' is compatible with conditions i-iv, so it is not clear why there is a need for a distinction in methodological type, but rather just a difference in details for each condition. For example, when a measurement is made, there is a (i) *physical interaction* that occurs, but it is broader than just the instrument and object. The interaction can include "experimental conditions" (Keyser 2017, 3-5). The product of a made measurement is also amenable to (ii) *theoretical characterization*. Keyser emphasizes that theoretical characterization is necessary for experiment/measurement (Keyser 2017, 14); but he does not make the additional move to say that theoretical characterization is *part of the process* of making a measurement. That is, in order to make a measurement about an effect, one needs to also *characterize* that effect. Without the final characterization, one is only dealing with the material conditions, which is an incomplete part of the measurement process. Keyser can accept that theoretical characterization is a necessary component of making a measurement. Otherwise, he risks offering a limited concept of 'making a measurement' that only applies to arranging the material components of the measurement process and nothing further.

The same challenge goes for (iii) *representational content* and (iv) *perspectival information*. An important component of the measurement process is to represent the relation between the produced effect and some aspect(s) of a phenomenon. For example, is this given effect a limited mimetic representation of a natural phenomenon or is it a brand new phenomenon? Without claims about what the effect is and its relation to objects, events, and processes in the world, 'making a measurement' is uninformative about part of the measurement process: the final value of the measurement outcome.

The aforementioned considerations question the need for a distinction between 'making' vs. 'taking'. One conclusion is that making uses the same components (i-iv), just with slightly different detail. But the other conclusion is a bit unsatisfying: making is really only about organizing the material components, which is an *initial* step in the measurement process, and it does not apply to later steps in measurement.

2.2. Dynamic relations between intervention and representation

I argue that the distinction between intervention vs. representation is less about *specific types of methodologies* in measurement/experiment and more about where to philosophically partition the *measurement process*. To make this point clear, I make two sub-points: 1) Measurement in the biological sciences offers complex and sometimes blurred relations between instrument and object of measurement such that representation and production take on dynamic roles; 2) There is a difference between the act of measurement and the total process of measurement. I briefly describe (1) and (2).

On van Fraassen's (2008) and Keyser's (2017) characterizations of *representation* in measurement, the role of the instrument/apparatus seems to have an important mediating function. It may be the case that philosophical focus on case studies (e.g., thermometry, microscopy, cellular bio, and bacteria) that are instrument-intensive provide a certain support for an instrument-centric account of representation in measurement. Whether or not the necessary mediating role of instruments is an explicit part of both accounts, there is room to develop a richer philosophical view of the role of representation in the total measurement *process*. Without such philosophical development, we risk missing complex cases of measurement where intervention occurs

side-by-side with representation. For example, in some cases of biological measurement, scientists use the organism to measure processes in that same organism but also to represent larger phenomena (Prasolova et al. 2006). For example, mouse diets are manipulated in order to measure chromatin pattern changes. I characterize this as the mouse *constituting experimental conditions* that are being manipulated in order to measure some sort of process. The manipulation of conditions indicates an interventionist approach (or 'making' a measurement). Moreover, without manipulating the mouse's diet scientists would not be able to make a reliable measurement on chromatin structure at all. So the organism is not only being manipulated as part of the experimental/measurement set-up, it is a crucial part of that set-up. That is, without intervention, there is no reliable result. In addition to the organism being used as part of the measurement set-up, it also serves as a physical *representation* of the dynamics of chromatin pattern change. That is, a given model organism can serve as a data model for a specific phenomenon of study—e.g., chromatin pattern in organism X. So, in this case the organism serves a dual function: it constitutes a set of experimental conditions to be manipulated and it serves as a physical representation of a phenomenon. Because of the dual function, this seems to be a case of both 'making' and 'taking' a measurement.

This brings me to sub-point (2). The total process of measurement is often complex in the biological sciences and requires multiple stages of intervening and representing. As mentioned in the model organism example representation and intervention are often *entangled*. Measurement is not merely putting an instrument up to something and waiting for a reading, which can be classified as an *act* of measurement. Measurement is also not merely creating effects out of material conditions. Measurement

Seattle, WA; 1-4 November 2018 -363-

requires manipulation of conditions that is *used* in order to generate a representation. For example, identifying a mysterious fungus that is entangled with other fungus in a sample is an active process that requires both intervention and representation. One method is to take a sample and scrape it over a petri dish. What grows are spores that are passively deposited. But if common fungi were commingled with the mysterious fungi in the sample, and the common fungi grew faster, it would be impossible to identify the mysterious fungus. That is, coming back in a couple of weeks and seeing the petri dish covered with familiar species would lead to a false conclusion. Another way to perform the measurement (i.e. culture samples) is as follows. Take the samples and grind them up. Then sprinkle them into a petri dish. Put the dish under the microscope and, using a fine needle, pick out fragments of the mysterious fungus and transplant them to their own dishes (Scott 2010). Once the fragments have been transplanted through this fine-grained intervention, each dish can be left to grow the colonies. The final dishes will offer visual representations that serve as data on the nature of the mysterious fungus. Notice here that intervention is a precursor to reliable representation.

Representation is not only reserved for the final instrument reading. It can also occur at other stages in the measurement process. Likewise, manipulation does not have to occur only at the earlier stages. For instance, organic matter can function as an instrument, like in the case of FourU thermometers, which are RNA molecules that act as thermometers in Salmonella (see Waldminghaus et al. 2007). Suppose that a scientist sets up an experiment to iteratively measure to what extent modifying RNA factors in FourU thermometers changes thermometer readings in Salmonella. In such a case the scientist could modify molecular factors and use the organic thermometers as temperature

measures over many iterations, which would culminate in some sort of data model that organizes the relationship between molecular factors and FourU function. In such a case, there are multiple layers of intervention and representation.

The complex layering of intervention and representation is apparent in biomarker measurement in the biomedical sciences, where biological components serve as representations of disease conditions, but are also intervened on in order to make more reliable representations. I turn to this case study in the subsequent section.

3. Intervening in Representations and Representing Interventions

Biomarkers are used in biomedical measurement to reliably predict causal information about patient outcomes while minimizing the complexity of measurement, resources, and invasiveness. A biomarker is an assayable metric—or simply, an indicator—that is used by scientists to draw conclusions about a biological process (De Gruttola et al. 2001). The greatest utility from biomarker measurement comes from their ability to help clinicians and researchers make conclusions with limited invasiveness. The reliance on biomarkers to make causal conclusions has prompted the use of 'surrogate markers'. These biomarkers are used to substitute for a clinically meaningful endpoint such as a disease condition. A major scientific methodological issue is that the use of multiple biomarkers will produce disagreeing results—and this is true even in the context of biomarkers that use similar biological pathways. To make methodological matters worse, theoretical representation is often not equipped to fill in the causal detail for each biomarker measurement. This amounts to an unfolding methodological puzzle about how to use intervention and representation in biomarkers to produce reliable measurements.

Seattle, WA; 1-4 November 2018 -365-

My interest in this case study is not in solving the methodological puzzle, but rather in showing the *relations between intervention and representation* in such a complex case study. In this section, I discuss the complexity of intervention and representation in biomarker measurement to illustrate how intervention mediates the measurement process.

To understand the complex methodology in biomarker measurement it is important to detail the use and limitations of biomarkers. Some biomarkers are used as a substitute for some clinical endpoint. For instance, LDL cholesterol (LDL-C) is a biomarker that clinicians and physicians use to correspond to a clinical endpoint—e.g., heart attack. Moreover, the biomarker is associated with risk factors such as coronary artery stenosis, atherosclerosis, and angina pectoris. Katz (2004) argues that all biomarkers are candidates for 'surrogate markers', which can serve as substitutes for clinical endpoints. That is, surrogate markers are reliable biomarkers that have a one-to-one correspondence with the disease condition such that they can be used to provide reliable predictive and causal information about a given clinical endpoint. There are a couple of points worth noting. First, notice that biomarkers and surrogate markers are being used as representations of a clinical endpoint. That is, to figure out the likelihood of developing a disease condition and to understand the risk factors associated with that disease condition, scientists use biomarkers that indicate information about the endpoint. This means that these physiological components can be used by clinicians and physicians to *represent disease conditions to respects and degrees*. The second point worth noting is that there are many biomarkers but limited surrogate markers and even more limited validated surrogate markers ('surrogate endpoints')—which are surrogate markers that are reliable in multiple contexts of interventions. The importance of this will be relevant

shortly when I discuss the complexity of biomarker measurement. For our purposes, this means that most biomarkers in biomedical practice provide very limited representational information.

Surrogate markers are not passively used as physical representations of disease conditions. Their use is often more effective for representational purposes if there is a *mediating intervention*. For instance, surrogate markers can constitute "response variables". This is where a surrogate marker is manipulated in order to produce an effect that is relevantly similar to the effect with the same manipulation on the clinical endpoint. This means that an adequate surrogate must be "tightly correlated" with the true clinical endpoint; but it also means that any intervention on a surrogate marker must be tightly correlated with the intervention on the true clinical endpoint (Buyse et al. 2000). I interpret this as a dual role for a reliable surrogate marker. It is to act as an epidemiological marker that *represents* some clinical endpoint but also to act as a responding variable that can be used in an *intervention* to causally influence the clinical endpoint. An example of the dual role of the surrogate marker is that high concentrations of LDL cholesterol (LDL-C) correspond to cardiovascular risk (Gofman and Lindgren 1950). But if a therapeutic interventions is used—such as, 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase inhibitors (statins)—that intervention can lower LDL levels, which in turn reduces cardiovascular disease (LaRosa et al. 2005).

So far I have presented the representational and intervention-based role of biomarkers. It is not straightforward to say that surrogate markers are *'made'* like an effect. But it is also not straightforward to say that surrogate markers constitute a *measurement outcome that is the final reading on an instrument*. These markers provide

Seattle, WA; 1-4 November 2018 -367-

useful representational information *in the context* of an intervention. To add to the complexity of the relation between representation and intervention, biomarkers in the context of Alzheimer's measurement have added methodological steps. In Alzheimer's measurement there are different biomarkers, which are not correlated with each other and change with independent dynamics in the progression of Alzheimer's disease. So *each* of these biomarkers do not provide the same type of representation about the progression of Alzheimer's disease. Furthermore, scientists *only* understand the disagreement between each of these biomarkers in the presence of different interventions.[1] The different interventions are in the form of drugs (e.g., bapineuzumab and solanezumab) and these interventions produce disagreeing representational results for the biomarkers. That is, the biomarkers respond differently to different interventions, which is methodologically problematic because it indicates that all of these biomarkers cannot be reliably tracking Alzheimer's progression in the same way. Interestingly, scientists systematically compare these disagreeing results to make reliable claims about Alzheimer's progression and treatment (Toyn 2015).[2] To simplify the method used, scientists track how interventions

[1] There has been much work recently on clinical biomarkers like: cerebrospinal fluid (CSF) tau, which is the primary component of neurofibrillary tangles; CSF 42-amino acid amyloid-β (CSF Aβ), which is the protein cleavage product believed to precipitate disease by forming neuron-damaging plaques; and amyloid plaques from PET scans. While the methodological story is beyond the scope of this discussion, there is a complex methodological point that is noteworthy for this discussion (Toyn 2015).

[2] To give a brief picture: The intervention of Bapineuzumab reduces levels of plaque assayed by Aβ PET and CSF tau, but not CSF Aβ; but Solanezumab *does not alter* levels

change properties of biomarkers and then they compare these amalgamated results with how interventions change behavioral/cognitive properties. This type of cross comparison allows scientists to eliminate biomarkers that do not track behavioral/cognitive improvement.

The structure of the methodological complexity in biomarker measurement can be partitioned as follows: 1) For a particular clinical endpoint, there are *limited physical representations* in the form biomarkers (or surrogate markers) which can be *used* to make representational and perspectival conclusions about the endpoint or risk factors associated with it; 2) *Scientists intervene in a process* from each of the biomarkers in order to track the relations between biomarkers and clinical endpoints; and 3) Such interventions prompt *disagreeing results between the biomarkers*, which can 4) be amalgamated by researchers into further representations of the *relations between biomarkers and their clinical endpoints*. The above structural breakdown is merely *a* type of complex methodological process that can occur in biomedical measurement. It shows how interventions on physical representations (biomarkers) can produce other reliable representations. What is important to note about this analysis is the role of intervention in *mediating* further representations. In the case of biomarkers, intervention is necessary to test how close biomarkers are in their representations of clinical endpoints and also to other biomarkers. These representations not only represent the relation between the original biomarker and the clinical endpoint, but they also represent how a given

---

of plaque assayed by Aβ PET and CSF tau but leads to a *reduction in* CSF Aβ. Cross comparison of the *intervention* mechanisms allows scientists to begin to make causal claims about which biomarkers are more reliable than others (Toyn 2015).

Seattle, WA; 1-4 November 2018 -369-

intervention affects a given biomarker. As such, intervention paves the way for iterations of representations.

4. Concluding Remarks

In this discussion, I have analyzed the role of intervention in mediating representations by using examples from the biological and biomedical sciences. Characterizing intervention as a mediating factor in a larger methodological operation provides an important point about scientific practice. Representation and intervention are not neatly partitioned into contrasting methodologies. In fact, applied science often dictates the complex, and often smeared, philosophical concepts and methodologies. For this reason, I am proposing a *process* view of intervention and representation. This view opens up the diversity of relations between representation and intervention in a given experimental/measurement practice. While I have emphasized how intervention mediates representation, there is more territory to explore about the mediating role of representation for intervention.

Work Cited

De Gruttola, V.G, Clax P, DeMets DL, et al. (2001). Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. Control Clin Trials 22:485–502.

Gofman, J.W., Jones, H.B., Lindgren, F.T., et al (1950). Blood lipids and human atherosclerosis. *Circulation* 2:161–178.

Hacking, I., (1983). *Representing and Intervening*, Cambridge: Cambridge University

Press.

Heidelberger, M. (2003). Theory-ladenness and scientific instruments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 138–151). Pittsburgh, PA: University of Pittsburgh Press.

Katz, R. (2004). Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* 1:189–195. doi: 10.1602/neurorx.1.2.189

Keyser, V. (2017). Experimental Effects and Causal Representations. *Synthese*, SI: Modeling and Representation, pp. 1-32.

LaRosa, J.C., Grundy, S.M., Waters, D.D., et al. (2005). Intensive Lipid Lowering with Atorvastatin in Patients with Stable Coronary Disease. *New England Journal of Medicine* 352:1425–1435. doi: 10.1056/NEJMoa050461

Prasolova L.A., L.N. Trut, I.N. Os'kina, R.G. Gulevich, I.Z. Pliusnina, E.B. Vsevolodov, I.F. Latypov. (2006). The effect of methyl-containing supplements during pregnancy on the phenotypic modification of offspring hair color in rats. *Genetika*, 42*(1)*, 78-83.

Radder, H. (2003). Technology and theory in experimental science. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 174–197). Pittsburgh, PA: University of Pittsburgh Press.

Toyn, J. (2015). What lessons can be learned from failed Alzheimer's disease trials? *Expert Rev Clin Pharmacol* 8:267–269. doi: 10.1586/17512433.2015.1034690

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.

Waldminghaus, T., Nadja H., Sabine B., and Franz N. (2007). FourU: A Novel Type of

RNA Thermometer in Salmonella. *Molecular Microbiology* 65 (2): 413–24. https://doi.org/10.1111/j.1365-2958.2007.05794.x.

Philosophy of Science (forthcoming)
v1.2 (as of 9/15/18)
Please cite published version

Are Emotions Psychological Constructions?

Charlie Kurth
Department of Philosophy
Western Michigan University

Abstract: According to psychological constructivism, emotions result from projecting folk emotion concepts onto felt affective episodes (e.g., Barrett 2017, LeDoux 2015, Russell 2004). Moreover, while constructivists acknowledge there's a biological dimension to emotion, they deny that emotions are (or involve) affect programs. So they also deny that emotions are natural kinds. However, the essential role constructivism gives to felt experience and folk concepts leads to an account that's extensionally inadequate and functionally inaccurate. Moreover, biologically-oriented proposals that reject these commitments are not similarly encumbered. Recognizing this has two implications: biological mechanisms are more central to emotion than constructivism allows, and the conclusion that emotions aren't natural kinds is premature.

This paper challenges the psychological constructivist account of emotions that is gaining prominence among neuroscientists and psychologists (e.g., Barrett 2017, 2012, 2009; LeDoux 2015; Russell 2004). According to constructivism, emotions result from projecting culturally-fashioned concepts onto felt affective episodes. Fear, for instance, just is a feeling of negative arousal as viewed through the lens of one's folk concept FEAR. This proposal is novel in taking felt experience and cognitive projection to be essential elements of what emotions are. Moreover, while constructivists acknowledge that there's a biological dimension to emotions (e.g., neural mechanisms are responsible for generating the conscious feelings that we project our emotion concepts on to), they deny that emotions are, or necessarily involve, anything like an affect program. Thus, constructivism is philosophically significant in two ways. First, in denying an essential role for biological mechanisms, it challenges influential, affect-program-oriented accounts of emotion (e.g., Scarantino & Griffiths 2011; Ekman & Cordaro 2011). Second, in understanding emotions as projections of folk emotion concepts, it takes emotions to be social-psychological constructions, not natural kinds.

Seattle, WA; 1-4 November 2018 -373-

But despite constructivism's appeal among cognitive scientists, the role that it gives to felt experience and folk concepts leads to an account of emotion that's both extensionally inadequate and functionally inaccurate. Moreover, biologically-oriented proposals that reject constructivism's problematic commitments are not similarly encumbered. Recognizing all this reveals that an adequate account needs to give greater place to the biological mechanisms that underlie emotions than constructivism allows. This, in turn, suggests that the constructivists' conclusion that emotions are not natural kinds is premature.

1. Psychological Constructivism and Its Appeal

Constructivism sees emotions as having two elements: a felt affective experience and a cognitive projection or labeling. Taking these in turn, the felt experience component—or "core affect" as it's often called—is a neurophysiological state that manifests as a consciously experienced combination of valence (i.e., feeling good or bad) and arousal (i.e., feeling activated or deactivated) (Barret 2006: 48; Russell 2004; LeDoux 2015: 226-232). Importantly, constructivism's focus on core affect looks just to the amalgamated *experience* of these two components—valence and arousal. What *causes* this felt experience is irrelevant to the nature and individuation of emotions. In fact, and as we will see, allowing that particular sensations (instances of core affect) can be produced by a range of distinct neural circuits or somatic events is taken to be a point in favor of the constructivist proposal.

Given this account of the felt dimension, constructivism maintains that "discrete emotions emerge from a conceptual analysis of core affect. Specifically, the experience of feeling an emotion…occurs when conceptual knowledge about emotion is brought to bear to categorize a momentary state of core affect. … [These] [c]ategorization processes enact the rules, [that guide] the emergence of an emotional episode" (Barrett 2006: 49; also LeDoux 2015: 225-232). This talk of "conceptual analysis," "conceptual knowledge," and "categorization" should be understood thinly.

The underlying process needn't involve some full-fledged, conscious judgment. Rather, all that's necessary is an unconscious or implicit recognition that one's sense of one's situation, and one's felt physiological state, fall under a particular folk emotion concept.

These emotions concepts, in turn, should be understood as folk theories or culturally-shaped behavioral scripts that detail the nature and function of the particular mental states picked out by specific emotion labels ('fear,' 'joy,' 'anger,' etc.). Moreover, the fact that folk emotion concepts engage these folk theories and behavioral scripts entails that the projecting of a particular label onto an instance of core affect not only imbues one's situation with the associated, emotionally-colored meaning, but also shapes one's subsequent thoughts, physiological responses, and behaviors (Barrett 2012; LeDoux 2015).

Formalizing this a bit, we can see psychological constructivism as committed to four theses:

> (PC1) Each emotion type/category is constituted by the projecting of a specific folk emotion concept (e.g., FEAR, JOY) onto a felt affective experience.
>
> (PC2) Token emotion episodes (e.g., a given instance of fear) are cognitive acts where one (implicitly) labels an occurrent conscious feeling with a particular folk emotion concept and so comes to see the feeling through the lens of that concept.
>
> (PC3) There is no unique (set of) neural circuit(s) or psychological mechanism(s) responsible for the conscious feelings that get categorized with particular folk emotion concepts.
>
> (PC4) The act of labeling a feeling with a particular folk emotion concept affects one's subsequent thoughts, physiological responses, and behaviors.

According to its advocates, much of constructivism's appeal lies in its explanatory power. In comparison to more biologically-oriented theories, it provides a better explanation of empirical research on the biological mechanisms and correlates associated with emotions (e.g., neural circuits, patterns of physiological change, and expressive behavior). Since the discussion that follows will build

Seattle, WA; 1-4 November 2018 -375-

from the contrast between constructivism and competing biologically-oriented theories (BTs), it will be useful to briefly sketch the BT approach and the constructivists' case against it.

As a generalization, BTs maintain that emotions are, or necessarily engage, affect programs—that is, largely encapsulated systems that automatically prompt stereotyped patterns of physiological changes, expressive behavior, motor routines, attentional shifts, and forms of higher-cognitive processing in response to (evolutionarily-relevant) threats and opportunities. So, for example, fear is (or essentially involves) an affect state that consists of automatically engaged tendencies for inter alia increases in arousal, narrowing of attention, and the cueing of fight/flight/freeze behavior in response to the perception of some danger.

But since BTs take affect programs to be essential (even identical) to emotions, constructivists argue they cannot explain two well-documented sets of findings.[1]

> (F1) One can feel a given emotion without engaging what science suggests is the best candidate for its underlying biological drivers (or their correlates)—e.g., activation of particular neural circuits, a distinctive physiological response, characteristic expressive behavior.
>
> (F2) The relevant biological drivers/correlates can be engaged though one does not report feeling the associated emotion.

So, for instance, though the central nucleus of the amygdala (CeA) is thought to be central to fear, research shows both that individuals will report being afraid when the CeA is not engaged (F1), and that the CeA can be active though individuals report not feeling fear (F2).

BT proponents have sought to address these explanatory limitations by insisting that we must narrow our understanding of what, say, FEAR is. More specifically, they maintain that the folk emotion concepts that the above research relies on (in, e.g., the self-reports of emotions (not) felt) are too

[1] See, e.g., Barrett 2012 for a review of the relevant empirical work.

coarsely grained for scientific investigations like these. The BT advocates' expectation is that a more refined account of what 'fear' refers to will reduce, even eliminate, dissociations of the sort noted above (e.g., Scarantino & Griffiths 2011; Kurth 2018). But constructivists respond that any effort to narrow or otherwise refine our emotion concepts along these lines will result in an account of (e.g.) fear that is troublingly stipulative or excessively revisionary with regard to our ordinary understanding of these emotions (Barrett 2012: 415-6; LeDoux 2015: 234).

Two aspects of this debates are particularly important for our purposes. First, central to the constructivist complaint is the move to take a failure to accommodate our *ordinary emotion talk* as the standard for what counts as stipulative or excessively revisionary account. Second, given our ordinary emotion talk as the standard, the above four theses appear to give constructivism the resources and flexibility it needs to explain not just (F1)-(F2), but also the richness and cultural variation of emotional life more generally (e.g., Barrett 2012, 2009). However, I will argue that investigating the extensional adequacy and functional accuracy of constructivism's core theses provides us with reason to doubt each of (PC1)-(PC4).

2. Is Constructivism Extensionally Adequate?

As we've seen, a central feature of the debate between constructivism and BTs is the charge that BTs cannot accommodate dissociation data without committing to a stipulative or excessively revisionary account of what emotions are. In what follows, I give three examples that suggest constructivism faces a similar problem. More specifically, a closer look at the constructivists' dual claim that emotions are *cognitive labelings* of *felt experiences* reveals that the account is both under- and over-inclusive with regard

Seattle, WA; 1-4 November 2018 -377-

to our ordinary understanding of things like: what emotions are, when we experience them, and how they differ from moods, feelings, and other categories of affect.[2]

First consider the constructivist's commitment to understanding emotions as felt experiences—that is, changes in core affect that we're consciously aware of. An implication of taking felt affective experience as essential to being an emotion is that it rules out the possibility of unconscious emotions. Some constructivists appear to embrace this result. For instance, Joseph LeDoux maintains that claims about unconscious emotions are "oxymoronic" (2015: 234; also, 19). But LeDoux's acceptance of this implication aside, the thought that there cannot be unconscious emotions fits poorly with our everyday experiences and our ordinary emotion talk.

For instance, if there aren't unconscious emotions, then how do we explain situations where we don't realize that we were (say) afraid until *after* the danger has passed? Pressing further, notice that we not only regularly speak of unconscious emotions, but also appeal to them in order to explain our behavior. For example, we say things like, "Bill won't discuss the book he is working on. He says it's not ready yet—but he doesn't realize that he's really just afraid about getting negative feedback." While ordinary talk like this is easy to make sense of on the assumption that Bill is unconsciously fearful, such an explanation isn't available to a constructivist like LeDoux—our ordinary talk to the contrary, Bill isn't unconsciously afraid, but rather experiencing some other psychological blockage.

But the constructivists' trouble with unconscious emotions runs deeper—the case for their existence also has empirical support. For instance, recent experimental work has shown that subliminally presented emotion faces can produce affective responses that bring emotion-specific behaviors *even though* the subject denies feeling an emotion. In particular, subliminally presented happy

[2] Thus the strategy I employ here—one that *grants* constructivists' their criterion for assessing when an account is excessively revisionary—is distinct from standard defenses of BTs noted in §1.

faces bring increased "liking" behavior (e.g., greater consumption of a novel beverage), while subliminally presented angry faces have the opposite result (Winkielman et al. 2003; also, Kihlstrom 1999). Since these patterns of behavior mesh with our understanding of both joy as an emotion that tends to increase interest/engagement, and anger as an emotion that brings avoidance/rejection tendencies, these results are taken as evidence of unconscious emotions.

While the constructivist might try to pass these findings off as cases where unconscious changes in core affect (not emotion) produce the behaviors, the plausibility of the proposal is undercut by the fit we find between the subliminally presented happy (angry) face, the resulting liking (avoidance) behavior, and *our ordinary understanding* what happiness (anger) involves (Winkielman et al. 2005). The upshot, then, is that constructivism's insistence that felt changes in core affect are *essential* to what emotions are has revisionary implications with regard to our ordinary (and scientific) understanding of emotional life.

But even if we're willing to grant that our talk of unconscious emotions is merely metaphorical—an elliptical way of talking about some non-emotion form of (unconscious) affect—the constructivist's second core commitment brings additional problems. In particular, the claim that emotions are the product of our cognitive labelings/projections makes facts about when we are experiencing an emotion—and what emotion it is—too sensitive to random situational features and framing effects. To draw this out, consider the following case.

> Coffee. I order a cup of decaf coffee and sit down to read a magazine cover story about Trump's latest foreign policy provocations. But unbeknownst to me, the barista confuses my order and I get a cup of regular coffee. As the caffeine works its way into my system, it brings a (consciously experienced) change in my arousal. As a result, I start reading the article with jittery attentiveness.

Given the scenario, it seems my jittery, attentive reading is best understood as a bout of caffeine-induced hyperactivity. But notice: there's nothing in the constructivist account to rule out the

possibility that I'm actually having an emotional experience—I'm afraid. After all, on the constructivist account, this experience could be a change in core affect that I've (implicitly) labeled 'fear.' While that possibility alone seems odd (to my ear, at least, the case is best understood as emotionless hyperactivity, not fear), there's more trouble.

To draw this out, consider the constructivist's likely response to the case. Given the setup, she would likely maintain that whether this is an instance of fear depends on whether I see it that way—what sort of meaning do I attribute to my situation (e.g., Barrett 2017: 126; 2012: 419-420; 2009: 1293)? For instance, if I assent to the barista's remark that I seem really uneasy about the article that I'm reading, then—by (implicitly) labeling my behavior through my assent—I imbue my situation with the meaning carried by my FEAR concept. I am, therefore, feeling fear. While this move might seem to allow the constructivist a way to account for the case, it comes at a high cost. For notice, had the barista instead said something like, "Whoops, I messed up and gave you regular, not decaf—no wonder you're so hyper," I'd likely assent to that too. And so I wouldn't be afraid—just hyperactively aroused.

But that's odd. Our ordinary thinking about emotions suggests that whether I'm experiencing a particular emotion, and what emotion I'm experiencing, should *not* be so sensitive to random situational features like what questions the barista—or anyone for that matter—just happen to ask me. To be clear, the claim here is not that emotions are immune to situational and contextual factors. Rather, the point is that on the constructivists' account emotions turn out to be *too* sensitive to them. The radical situational sensitivity entailed by constructivism makes it not only too easy to experience an emotion, but also ties facts about what emotion we're experiencing to irrelevant situational factors.

Together, the difficulties raised by unconscious emotions and incidental situational features call the extensional adequacy of the constructivist account into question and do so in a way that

pinpoints the commitments of (PC1) and (PC2) as the source of the trouble—after all, these claims posit feelings of core affect and projections of folk concepts as essential to what emotions are. Of equal note is the fact that biological theories are less vulnerable to these difficulties. For one, irrelevant situational features should have less influence on what emotion one happens to experience since, according to BTs, emotions are (or are principally driven by) affect programs, not contextualized cognitive labelings. Moreover, since affect programs are things that can operate below that level of conscious awareness (Kurth 2018), taking emotions to be driven by affect programs provides BTs with the resources needed to explain unconscious emotions.

While the above discussion raises worries about the first two constructivist theses (PC1-PC2), it also provides the makings for worries about the third. In particular, because constructivism denies (via PC3) that emotions are underwritten by affect programs, it has trouble making plausible distinctions between emotions and similar states like moods. To draw this out, notice that the coffee case from above can be easily extended to show that constructivism makes it too easy to flip between moods and emotions. All we need to do is substitute "being in a worried mood" for "hyperactive" in the presentation of the case. Once we do this, we see that mere changes in the question the barista asks me can change whether I'm worried (a mood) or afraid (an emotion).

So we again see that constructivism has problematic explanatory limitations—this time with regard to preserving the thought that there's a substantive difference between moods and emotions. On the constructivist account, this distinction is just a matter of how we happen to label our felt experiences. While some constructivists appear willing to accept this conclusion (e.g., Barrett 2017, 2009), it highlights another place where the constructivist proposal has revisionary implications—after all, moods and emotions are generally thought to be *distinct* forms of affect (e.g., Ben-Ze'ev 2000: Chap. 4). Moreover, here too we have a difficulty that's easily avoided by biological accounts. Since

Seattle, WA; 1-4 November 2018 -381-

BTs take emotions to be (driven by) affect programs, they can appeal to the engagement of these mechanisms as the basis for the emotion/mood distinction (e.g., Kurth 2018; Wong 2017).

Stepping back, then, although constructivism purports to be less stipulative with regard to capturing our ordinary understanding of emotions, the above examples call this into question. For starters, the constructivists' commitment to (PC1)-(PC3) has revisionary implications for our ordinary understanding of what emotions are, when we experience them, and how they differ from moods. Moreover, we have also seen that biologically-oriented accounts—in eschewing this trio of problematic theses—are better equipped to provide a plausible account of these features of our everyday emotion talk.

3. Is Constructivism Functionally Accurate?

The challenges to the constructivist picture extend beyond concerns about its extensional adequacy. The account also makes predictions about how projecting emotion concepts onto felt experience should shape subsequent behavior that are poorly supported by the empirical record. Two examples will draw this out.

First consider emotion misattribution research. In this work, a feeling that is typically associated with a particular emotion (e.g., feelings of unease and anxiety) is subtly induced, but the individual is lead to believe they are not, in fact, experiencing that emotion but rather something else (e.g., the effects of caffeine). Constructivism predicts (via PC4) that individuals in these experiments should display different behaviors depending on whether they are in the control or misattribution conditions. For instance, individuals led to believe that the unease they're feeling is not anxiety, but something else (caffeine) should display diminished anxiety-related behaviors in comparison to controls who were not misled about their unease. But on this score, the experimental findings are decidedly mixed.

First, while there is a sizable body of findings showing misattribution manipulations attenuate subsequent emotion-related behavior, there is also a sufficiently large set of non-confirmations to raise concerns. For instance, while some research on public speaking anxiety suggests that attributing unease to a pill you just took rather than anxiety about a public talk you must give leads to a reduction in anxiety-related behaviors—stuttering, apprehension, and the like (Olson 1988), other studies have failed to find any differences in these behaviors (Slivkin & Buss 1984; Singerman, Borkovec & Baron 1976).

Moreover, even in cases where emotion-related behavior is reduced in the manipulation condition, it's not clear how much support this brings to the constructivist. This is because it's often unclear whether the reductions in emotion-specific behavior are (i) the result of the misattribution or (ii) a consequence of directing subjects' attention away from the emotion eliciting stimuli (for a review, see, e.g., Reisenzein 1983). This potential confound is problematic for constructivists since only possibility (i) provides direct support for the claim of (PC4)—namely, that the act of labeling *itself* affects subsequent behavior.

The second problematic set of results comes from work in political science. This research investigates how negative emotions shape public policy decision making among voters (e.g., MacKuen et al. 2010; Brader et al. 2008; Valentino et al. 2008). The core hypothesis of this research is that negative emotions (especially, anger and anxiety) affect subsequent behavior in different ways. In particular, anger—as a response to challenges to what one values—should tend to bring behavior geared toward defending the threatened values. By contrast, since anxiety is a response to uncertainty, it should tend to bring caution and information gathering aimed helping one work through the uncertainty one faces.

Seattle, WA; 1-4 November 2018 -383-

To test these predictions, the experimental set up works as follows. First, individuals are asked to read a (fake) news story designed to provoke anger or anxiety by challenging the individuals' pre-existing views about contentious policy issues like immigration, affirmative action, and economic policy. After reading the story, the participants are given the opportunity to use a website containing links to additional information, both for and against, the policy issue at hand. They are also asked how the original news story they read made them feel (e.g., angry, anxious). So by tracking what kinds of information the participants looked at through the website, experimenters can identify differences in how the anger and anxiety provoked by the story shaped subsequent behavior.

In the present context, these experiments allow us to test a pair of predictions that follow from the constructivist theses (PC1) and (PC4):

> (P1) Labeling felt experiences with distinct folk emotion concepts should bring different patterns of behavior.
>
> (P2) The behaviors that result from labeling a felt experience with a particular concept should map to our folk understanding of the emotion in question.[3]

More specifically, given (P1) and (P2), we should see different behaviors based on whether the participants in the experiment label their emotion 'anger' or 'anxiety' (P1). Moreover, the different behaviors should map to the above, ordinary understanding of these emotions—e.g., angry individuals should look for information that helps them defend their preferred policy position, while anxious individuals should engage less in motivated inquiry and more in open-minded forms of investigation (P2).

[3] As evidence of constructivism's commitment to these predictions, consider Lisa Feldman Barrett's comment that "when a person is feeling angry…she has categorized sensations from the body and the world using conceptual knowledge of the category 'anger'. As a result, that person will experience an unpleasant, high arousal state as evidence that someone is offensive. In fear…she will experience the same state as evidence that the world is threatening. And, *either way, the person will behave accordingly*" (2009: 1293, emphasis added).

However, whether we find support for these predictions turns—surprisingly—on what the policy issue used in the experiment was. More specifically, in experiments where the policy question that was challenged by the fake news story concerned immigration, the results fit poorly with constructivism's predictions. That is, participants behaved in the same angry way regardless of whether they reported feeling anger or anxiety (Brader et al. 2008). By contrast, if the policy issue at hand concerned affirmative action or economic policy, the results are more in line with (P1)-(P2): anger and anxiety provoked by the news stories not only brought different patterns of behavior, but the resulting behaviors mesh with our ordinary conception of how these emotions function (MacKuen et al. 2010; Valentino 2008).

While this second set of results might appear to be good news for constructivists, the trouble lies in explaining why we get the different results between the immigration and affirmative action/economic policy experiments. After all, other than the content of the issue at hand, the experimental designs were *identical*. In response, the constructivist might argue that content and context matter (e.g., Barrett 2012, 2009): the similar behaviors that subjects display in the immigration version of the study suggest that the cultural scripts associated with 'anger' and 'anxiety' are highly sensitive to negative stereotypes about minorities. More specifically, the thought would be that there's something about the combination of immigration debates and racial stereotypes that changes the standard behavioral scripts associated with 'anger' and 'anxiety' so that, while they *typically* generate different behaviors, they *now* bring the same ones.

But setting aside concerns about the ad hoc nature of this proposal, without more of a backstory, it's unconvincing. After all, affirmative action debates are *also* framed in racial stereotype provoking ways. So here too we should see anger and anxiety generating similar patterns of behavior. But we don't.

Seattle, WA; 1-4 November 2018 -385-

Moreover, notice that, on this front, biological accounts have an easier time explaining the experimental findings. For instance, as one possibility, the BT advocate could argue that only participants in the immigration study are likely to be experiencing *both* anger and anxiety: anger about the harms immigrants will bring and anxiety given their uncertainty about the likelihood of these harms. Given this, the BT advocate could then add two claims about what happens when both these emotions are engaged. First, since anger is a more powerful emotion than anxiety, it tends to win out with regard to shaping individuals' subsequent behavior. Second, given the high degree of overlap in the felt experiences produced by the anger and anxiety affect programs (e.g., both bring increased, negatively valenced arousal), when prompted to state what emotion they are feeling, some subjects happen to interpret their feelings as anger, while others see it as anxiety. Thus, the BT advocate can explain both why we get mixed results when subjects are prompted to state what emotion they are feeling and why, despite these differences in self-reports, the individuals nonetheless respond with behavior characteristic of anger, not anxiety. Moreover, because this proposal allows anger to drive behavior *regardless* of how subjects happen to label it, the explanation is unavailable to constructivists.

All told, we have two independent sets of experimental findings showing (at best) equivocal support for constructivism's predictions about how projecting emotion concepts onto felt experience should shape subsequent behavior. Moreover, we've also learned that more biologically-oriented accounts are better able to handle the experimental findings we've reviewed.

4. Conclusion: Emotions, Biology, and Natural Kinds

As we've seen, constructivism's purported advantage over more biologically-oriented theories lies is its ability to better explain the richness and diversity of emotional life (§1). But we have also seen that a crucial premise in this argument is the move to take accommodating our ordinary emotion talk as the standard for assessing a theory's explanatory power. Not only are there familiar problems for

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -386-

adopting such a standard (e.g., Scarantino & Griffiths 2011, Kurth 2018), but—even if we accept it—we've learned that there's trouble for constructivism. In particular, the explanatory "success" constructivism secures come by way of a highly revisionary account of what emotions are, when we experience them, how they differ from moods, and the way that they shape behavior (§§2-3). Moreover, our critical observations also implicate the four constructivist theses (PC1-PC4) as the source of these difficulties. Thus it's not surprising that more biologically oriented proposals—accounts that reject these commitments—do not face similar explanatory limitations.

Taken together, then, the arguments of this paper suggest a pair of larger lessons. First, even if we agree that constructivists are correct about what the relevant standard for assessing a theory of emotion is, we've learned that an adequate account must give greater place to the biological mechanisms that underlie emotions than constructivism allows. This, in turn, indicates that the constructivists' conclusion that emotions are not natural kinds is premature. After all, if we must posit something like an affect program in order to (i) explain everyday talk and empirical findings about unconscious emotions, (ii) capture the thought that emotional experience is not radically sensitive to random situational features, and (iii) accommodate research regarding how emotions shape behavior, then we have evidence that (at least some) emotions are underwritten by mechanisms that make them plausible candidates for being natural kinds.

References

Barrett, L. 2017. *How Emotions Are Made*. New York: Houghton Mifflin Harcourt.

---. 2012. "Emotions Are Real." *Emotion* 12: 413-429.

---. 2009. "Variety is the Spice of Life." *Emotion and Cognition* 23: 1284-1306.

---. 2006. "Emotions as Natural Kinds?" *Perspectives on Psychological Science* 1: 28-58.

Ben-Ze'ev, A. 2000. *The Subtlety of Emotions*. Cambridge.

Brader, T. et al. 2008. "What Triggers Public Opposition to Immigration?" *American Journal of Political Science* 52: 959-978.

Ekman, P. & D. Cordaro. 2011. "What is Meant by Calling Emotions Basic." *Emotion Review* 3: 364–370

Kihlstrom, J.F. 1999. "The Psychological Unconscious." In L.A. Pervin & O.P. John (Eds.), *Handbook of Personality* (2nd ed., pp.424–442). New York: Guilford Press.

Kurth, C. 2018. *The Anxious Mind*. MIT Press.

LeDoux, J. 2015. *Anxious*. New York: Viking.

MacKuen, M. et al. 2010. "Civil Engagements," *American Journal of Political Science* 54: 440–458.

Olson, J. 1988. "Missattribution, Preparatory Information, and Speech Anxiety" *Journal of Personality and Social Psychology* 54: 758-767.

Reisenzein, R. 1983. "The Schachter Theory of Emotion" *Psychological Bulletin* 94: 239-264.

Russell, P. 2004. "Core Affect and the Psychological Construction of Emotion. *Psychological Review* 110: 145–172

Scarantino, A & P. Griffiths. 2011. "Don't Give Up on Basic Emotions" Emotion Review 3: 1-11.

Singerman, K. et al. 1976. "Failure of a 'Misattribution Therapy' Manipulation with a Clinically Relevant Target Behavior" *Behavior Therapy* 7: 306-316.

Slivken, K. & A. H. Buss. 1984. "Misattribution and Speech Anxiety" *Journal of Personality and Social Psychology* 47: 396-402.

Valentino, N. et al. 2008. "Is a Worried Citizen a Good Citizen?" *Political Psychology* 29: 247–73.

Winkielman, P. et al. 2005. "Unconscious Affective Reactions to Masked Happy versus Angry Faces Influence Consumption Behavior and Judgments of Value." *Personality and Social Psychology Bulletin* 121-135.

Wong, M. 2017. "The Mood-Emotion Loop" *Philosophical Studies* 173: 3061-3080.

Symposium: *Bridging the Gap Between Scientists and the Public*, PSA 2018

How trustworthy and authoritative is scientific input into public policy deliberations?[i]

Hugh Lacey
Swarthmore College / University of São Paulo

Abstract: Appraising public policies about using technoscientific innovations requires attending to the values reflected in the interests expected to be served by them. It also requires addressing questions about the efficacy of using the innovations, and about whether or not using them may occasion harmful effects (risks); moreover, judgments about these matters should be soundly backed by empirical evidence. Clearly, then, scientists have an important role to play in formulating and appraising these public policies.

However, ethical and social values affect decisions made about the criteria (1) for identifying the range of risks, and of relevant empirical data needed for making judgments about them, that should be considered in public policy deliberations, and (2) for determining how well claims concerning risks should be supported by the available data in order to warrant that they have a decisive role in the deliberations. Consider the case of public policies about using GMOs. Concerning the range of data: is it sufficient for risk assessment only to be informed by data relevant to investigating the risks of using GMOs that may be occasioned by way of physical/chemical/biological mechanisms directly triggered by events within their modified genomes? Or: should data pertaining to the full range of ecological and socioeconomic effects of using them, in the environments in which they are used and under the socioeconomic conditions of their use, also inform this assessment? Those interested in producing and using GMOs, in the light of their adhering to values of capital and the market, are likely to give a positive answer to the first question; those holding competing values, e.g., connected with respect for human rights and environmental sustainability, to the second. And, concerning the degree of support: the former – citing the ethical gravity of losses (both economic and, allegedly, for food security) that would be incurred by failing to use GMOs on a wide scale – are likely to require less stringent standards of evidential appraisal than the latter.

Scientists, *qua* scientists, however, do not have special authority in the realm of values. Thus, their judgments, about the evidential support that claims about risks (and some other matters) have, may sometimes be reasonably (although not decisively) contested partly on value-laden grounds – as they have been in the GMO case, where the contestation has generated considerable controversy, and continues to do so. It follows that, in the context of deliberations about public policy, unless scientists engage with representatives of all stakeholders in the outcomes of the policies (as, for the most part, has not happened in the GMO case) – taking into account that their competing values may lead to making different decisions about what are the relevant data, as well as about the degree of support required for their claims about risks to gain the required credibility to inform the deliberations; and respecting "tempered equality" of participants in the dialogue (Longino) – their trustworthiness is put into question and their authority diminished.

1.

In a letter, dated June 29th, 2016, 135 Nobel laureates made the following claims, among others,[ii] related to using GMOs (genetically modified organisms) in agriculture:

(i) "Scientific and regulatory agencies around the world have repeatedly and consistently found crops and foods improved through biotechnology to be as safe as, if not safer than those derived from any other method of production."

(ii) "There has never been a single confirmed case of a negative health outcome for humans or animals from their consumption."

(iii) "Their environmental impacts have been shown repeatedly to be less damaging to the environment, and a boon to global biodiversity" (Laureates Letter, 2016).

Reflecting the authority and esteem that tends to be accorded to Nobel laureates, the declaration was widely reported and taken to bolster the allegation that there is a *scientific consensus* that cultivating and harvesting genetically engineered crops, and consuming their products, is safe.[iii] The scientists who signed it aimed to assure the public that the three claims are well con-

Seattle, WA; 1-4 November 2018 -389-

firmed, and that public policy and regulatory deliberations should reflect them. The claims do not derive from outcomes of the research conducted by these scientists, for at most one or two of them (so far as I can tell, none) have themselves engaged in biosafety research. They were putting their authority behind the research and judgments of others, whom presumably they trusted. Even so, one might reasonable assume that they had, before signing the declaration, examined the relevant research and concurred with its outcomes, and had found good reason to tell us, as they do, (presumably based on a thorough examination of its writings and actions) that the opposition is "based on emotion and dogma contradicted by data" and that it "must be stopped." At the end of the paper, I will argue that the declaration misuses scientific authority and contributes to doubts about the trustworthiness of leading scientific authorities. My larger purpose, however, is to suggest **some** necessary conditions for re-establishing trust in scientific communities – bridging the gap between scientists and the public, and (the concern of de Martín-Melo & Intemann, 2018) – so that both the authority and integrity of science, and the conditions for strengthening democratic societies, are enhanced

2.

First, some more general remarks. I maintain that the deliberations out of which arise public polices having to do with introducing, using and regulating technoscientific innovations (I only have time to discuss GEOs) should consider:

(1) questions about the *efficacy* of the proposed uses are addressed – and about their *safety*, specifically about how well available empirical evidence confirms that the proposed uses do not occasion harmful effects (or risks of causing harmful effects);

(2) the values reflected in the interests expected to be served by the proposed uses, as well as questions about whether interests expected to be served by competing values may be disadvantaged by them, and priorities among the competing interests;

(3) identified potential alternatives to using these innovations – including fundamentally different kinds of practices – as well as how using them compares to the proposed uses with respect to efficacy and safety (and other potential benefits).[iv]

Of these conditions only (1) is uncontroversial and generally followed (although there are disagreements about how it ought to be followed) in public policy deliberations.[v] Clearly satisfactory answers to the questions about efficacy and safety depend on trustworthy and reliable scientific input. I will not question that scientific research has reliably established the efficacy of the GEOs that have already been approved by regulatory bodies for agricultural use, for the most part GEOs with herbicide-resistant and insecticidal properties.[vi] Efficacy does not imply safety, however, and the research approaches (in molecular biology, biotechnology, etc) within which efficacy is established do not suffice for engaging in research dealing with safety. However, many regulatory practices presuppose that scientific input, pertaining to deliberations about safety – like that about efficacy – is obtained prior to consideration of (2) and (3), and to entanglement with value questions. Hence, the currency of the terms "scientific risk assessments" and

"scientific safety studies", areas of research in which scientific/technical "experts" should be granted authority.

One needs to be wary here, for "safe" and "risk" are 'thick ethical terms'. Scientific safety studies cannot be fully separate from entanglement with values and obligations. Thus, e.g. (simplifying a little), 'using X *is* unsafe' implies (*ceteris paribus*) 'X *should* not be used, unless appropriate precautions are taken.' And, when scientists conclude, on the basis of their investigations, that 'using X is safe', they intend it to follow (and to have impact at step (2)), that *ceteris paribus* 'it is improper to impede using X'.[vii] This does not mean that, in the course of empirical research in scientific safety studies, value-laden terms are used in articulating hypotheses and reporting empirical data. The link between the results of the empirical research and the subsequent value judgments depends on a step (call it step (0)), casually made prior to the empirical investigations. At step (0), the set of possible unintended collateral effects of using X is scrutinized, and those that are identified as harmful (as risks)[viii] – obviously value judgments are made here – are then investigated for such matters as the probability and magnitude of their possible occurrence, and its being countered by introducing scientifically informed regulations. In the investigation, the possible collateral effects are characterized, not with thick ethical terms, but with theoretical and observational terms deployed in relevant scientific fields, like molecular biology, chemistry, soil sciences and physiology (whose terms have no value connotations). Then, 'using X is safe' may be concluded,[ix] – usually qualified by 'provided that it is used in accordance with stipulated regulations' – if the investigations confirm that none of the investigated effects would occur with significant magnitude and probability when X is used in accordance with the regulations. This account is consistent with the picture of scientific safety studies that has step (1) preceding steps (2) and (3); but it clarifies that the move from empirically confirmed results at (1) to the claim the value-implicated 'X is safe' and to value judgments of relevance at (2) rests upon value judgments made at step (0). It follows that the conclusion, 'X is safe', might appropriately be challenged – without thereby challenging the scientists' judgments about each of the particular possible effects investigated – on the basis of the value judgment that not all the harmful possible effects of using X were identified at (0).

The outcomes of "scientific" safety studies usually constitute the only input to the deliberations of the 'technical' commissions that participate in public policy deliberations about using and regulating technoscientific objects. In these studies (in the GEO case), at step (0), the possible effects identified as harmful are a subset of those that may be occasioned by way of physical/chemical/biological mechanisms directly triggered by events within the modified genomes of plants. One can identify *two ways in which the adequacy of these studies might be challenged*.[x]

First: Conclusions drawn about the safety of using V (a genetically engineered plant variety) could be challenged on the ground that the subset chosen for investigation does not include some possible effects, with similar mechanisms, that are of of special salience for those who uphold a particular value-outlook.[xi] For them, even well conducted studies on the the items of the subset chosen will be insufficient to confirm that using V is safe.[xii] Challenges of this type can be

Seattle, WA; 1-4 November 2018 -391-

resolved (in principle) by conducting more scientific studies of the same kind after having identified a larger relevant subset.[xiii]

Second: Their adequacy could be challenged by those, who object that the set from which the subsets are chosen for "scientific safety studies" is not sufficiently encompassing. For them, deliberations about the safety of using GE-plants should be informed by appropriate empirical investigations, not only of potential effects occasioned by way of physical/chemical/biological mechanisms directly triggered by events within their modified genomes, but also the full range of potential ecological and socioeconomic effects occasioned by using them in the environments (agroecosystems) of their actual or intended use, and under the socioeconomic conditions of their use, taking fully into account that the potential effects vary from variety to variety and species to plant species. Upholding values of respect for human rights, democratic participation and environmental sustainability, which are opposed to those of capital and the market, often motivates challenges of this kind. These potential effects cannot *all* be investigated in "scientific safety studies," for they require utilizing ecological, human and social categories that have no place in research in such areas as physics, chemistry, and molecular biology, and that may include thick ethical terms (e.g., food security, being poisoned).[xiv] To investigate them empirically, therefore, requires adopting methodological approaches that are not reducible to those used in the indicated scientific areas, and that are generally outside of the expertise of scientists trained in the methodologies appropriate to them. The expertise required to engage in research that leads to the development of GEOs is quite different from that required for studies about the safety of using them.

At issue here are not only concerns about risks (potential harmful effects). Farmers (and their communities) in many areas of the world have suffered serious health problems because of having been exposed to glyphosate (the principal active ingredient in the widely used herbicide, RoundUp) sprayed on fields planted with glyphosate-resistant GEOs.[xv] They are unimpressed when told that the varieties of GEOs planted in these fields had undergone and passed "scientific safety tests." They know from their experience (even if it is not well recorded in peer reviewed studies) that, regardless of what was the case in the conditions of the tests, it is not safe to cultivate these GEOs (which require the accompanying use of glyphosate) in the ways and under the conditions in which they are used in their locales. And, they continue to be unimpressed when the manufactures and regulators of the GEOs insist that the problem was not with cultivating the GEOs, but with using glyphosate without heed to stipulated regulations for safe use,[xvi] for they have good reason to believe that the sellers of GEOs and glyphosate know that they will in fact not be used in accordance with these regulations.[xvii]

3.

Summing up, ethical and social values properly affect decisions (at step 0)) made about the criteria to be deployed for identifying the range of risks that should be considered in public policy deliberations, and of the relevant kinds of empirical data needed for making judgments about them. They also – consistent with maintaining that judgments about safety (step (1)) can be settled

prior to steps (2) and (3) – also affect the standards deployed for determining how well claims about risks should be supported (by the available empirical data) – in order to ensure that risks are dealt with properly in public policy deliberations.

Those who uphold values of capital and the market (agribusiness corporations, governments that prioritize economic growth, etc) are likely to cite the ethical gravity of losses (both economic and, allegedly, for food security) that would be incurred by failing to use GEOs on a wide scale; and consequently to require less stringent standards of evidential appraisal than those who uphold values of respect for human rights, democratic participation and environmental sustainability, who are likely to adopt precautionary stances that permit time for research incorporating more stringent standards to be met.[xviii] Similarly, those who uphold the latter values are likely to emphasize the importance of step (3): investigating alternatives to the food/agricultural system, in which using GEOs and the use of agrotoxics are acquiring ever larger roles, alternatives such as agroecology, a scientifically-informed approach to agriculture that attends simultaneously to production, sustainability, social health, strengthening the values and cultures of local communities, and to furthering the practices needed to implement policies of food sovereignty – and to urge the public support of research, in which are adopted strategies appropriate for dealing with the human, ecological and social dimensions of agroecosystems.[xix]

Scientists, *qua* scientists, however, do not have authority in the realm of ethical and social values. The values they uphold, even when widely shared, do not trump those upheld by other groups in democratic public policy deliberations. Thus, their judgments, about the evidential support that claims about the safety of planting GEO crops and consuming their products have, may sometimes be reasonably contested partly on value-laden grounds (cf. de Melo-Martín & Intemann, 2017, p. 131). That contestation cannot be rebutted by appeal to the alleged "scientific consensus" that GEOs (or, particular varieties of them) are safe. Apart from the fact that actually there is no such consensus, manifestly so among experts in biosafety investigations,[xx] if there were, it would likely secrete the scientists' shared value commitments, a matter on which they have no authority. Appeal to such an alleged consensus covers up the role of upholding the values of capital and the market in affirming it.

It follows that, in the context of deliberations about public policy, the trustworthiness of scientists is put into question and their authority unmerited,

- unless they engage with representatives of all stakeholders in the outcomes of the policies (as, for the most part, has not happened in the GEO case);
- unless, moreover, in doing so – respecting what Longino (2002, p. 129–135) calls "tempered equality" of participants in the deliberations – , they take into account that upholding competing values (e.g., of company-employed scientists and family farmers) may lead to making different judgments concerning relevant data, hypotheses to investigate, and approaches to farming, as well as concerning the degree of support required for claims about safety to merit credibility.

4.

5

Seattle, WA; 1-4 November 2018 -393-

Let us now return to the three claims (introduced at the outset) that the 135 Nobel laureates endorsed:[xxi]

These claims are ambiguous, misleading, in some instances false, and apparently made without acquaintance with the relevant studies and arguments of their critics. (i) is false: I am not aware of any agency that has compared the safety of GEO crops and their food products with that of agroecological (or organic farming) methods of production – the agencies have not sought out the results of research dealing with that comparison (and very little of it has been conducted). At most, they have found GEO crops and products to be at least as safe as conventional high-input crops and their products, but that doesn't respond to the critics who endorse agroecological methods of production. (ii) is probably true – but misleading: it does not mention that epidemiological studies of consumption of GEOs have not been conducted,[xxii] to a large extent because legal prohibition of labelling GEO products poses probably an insurmountable impediment to conducting them; and that it is well documented that cultivating GEOs has occasioned health problems for numerous farmers who have been exposed to the agrotoxics, whose use is integral to the cultivation of certain varieties of GEOs. (iii) is ambiguous: the environmental impacts may indeed be less damaging than those of conventional high-input agriculture; but they are incomparably more damaging to the environment than agroecological farming that has environmental sustainability built into its fundamental objectives.

By dismissing criticisms like these "based on emotion and dogma contradicted by data," and not attempting to rebut them in a context where something like Longino's conditions are in place, the scientists undermine the authority that science should be able to demand to be recognized; and they weaken the contribution that science could make to democratic policy deliberations.

References

Bombardi, Larissa M. (2017) Geografia do Uso de Agrotóxicos no Brasil e Conexões com a União Europeia. E-book, <https://drive.google.com/file/d/1ci7nzJPm_J6XYNkdv_rt-nbFmOETH80G/view>. São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

De Melo-Martín, I. and Intemann, K. (2018) *The Fight against Doubt: How to bridge the gap between scientists and the Public*. New York: Oxford University Press.

Hilbeck, A., Binimelis, R., Defarge, N., Steinbrecher, R., Székács, A., Wickson, F., Antoniou, M., Bereano, P. L., Clark, E. A., Hansen, M., Novotny, E., Heinemann, J., Meyer, H., Shiva, V. & Wynne, B. (2015) No scientific consensus on GEO safety. *Environmental Sciences Europe* 27: 4–9.

Human Rights Watch (2018) The Failing Response to pesticide Drift in Brazil's Rural Communities, July 20, 2018, <https://www.hrw.org/report/2018/07/20/you-dont-want-breathe-poison-anymore/failing-response-pesticide-drift-brazils>.

Krimsky, S. (2015) An illusory consensus behind GEO health assessment. *Science, Technology and Human Values* 40 (6): 883–914.

Lacey, H. (2005) *Values and Objectivity in Science; current controversy about transgenic crops.* Lanham, MD: Lexington Books.

— (2015a) Food and agricultural systems for the future: science, emancipation and human flourishing. *Journal of Critical Realism* 14 (3), 2015: 272–286.

— (2015b) Agroécologie : la science et les valeurs de la justice sociale, de la démocratie et de la durabilité. *Ecologie et Politique*, No. 51, 2015: 27–40.

— (2016) Science, respect for nature, and human well-being: democratic values and the responsibilities of scientists today. *Foundations of Science* 21(1): 883–914.

— (2017) The safety of using genetically engineered organism: empirical evidence and value judgments. *Public Affairs Quarterly* 31 (4): 259–279.

Lacey, H., Corrêa Leite, J., Oliveira, M.B., & Mariconda, P.r. (2015a) Transgênicos: malefícios, invasões e diálogo. *JC Notícias*, Edition 5167 (April 30, /2015), http://www.jornaldaciencia.org.br/edicoes/?url=http://jcnoticias.jornaldaciencia.org.br/9-transgenicos-maleficios-invasoes-e-dialogo/.

— (2015b) Transgênicos: diálogo. *JC Notícias*, Edition 5182 (May 22/2015), http://www.jornaldaciencia.org.br/edicoes/?url=http://jcnoticias.jornaldaciencia.org.br/27-transgenicos-dialogo/.

Longino, H. (2002) *The Fate of Knowledge*. Princeton: Princeton University Press.

Laureates Letter (2016) "Laureates letter supporting precision agriculture," http://supportprecisionagriculture.org/nobel-laureate-gmo-letter_rjr.html .

US National Academies of Science, Engineering and Medicine (2017). *Genetically Engineered Crops: Experiences and Prospects*. Washington: National Academies Press.

Paganelli, A., Gnazzo, V, Acosta, H., López, S.L. & Carrasco, A.E. (2010) 'Glyphosate-based herbicides produce teratogenic effects on vertebrates by impairing retinoic acid signaling'. *Chemical Research in Toxicology* 23: 1586–1595.

Traavik, T. & Ching, L.L. (2007) *Biosafety first: Holistic approaches to risk and uncertainty in genetic engineering and genetically modified organisms*. Trondheim, Norway: Tapir Academic Press.

Appendix

The central concern of the letter signed by the Nobel laureates is to support the program of research on Golden Rice [a variety of genetically engineered rice] and to denounce opposition to it, especially that of the NGO, Greenpeace. In a longer work, I would also discuss critically the way in which the letter misleads both about the state of research on Golden Rice and about that character of criticisms that question the importance of this research.

(a) The letter states that Greenpeace "has spearheaded opposition to Golden Rice, which has the potential to reduce or eliminate much of the death and disease caused by a vitamin A deficiency, which has the greatest impact on the poorest people in Africa and Southeast Asia". It called upon "governments of the world to reject Greenpeace's campaign against Golden Rice specifically, and crops and foods improved through biotechnology in general; and to do everything in their power to oppose Greenpeace's actions and accelerate the access of farmers to all the tools of modern biology, especially seeds improved through biotechnology"; and concluded with the warning: "Opposition based on emotion and dogma contradicted by data must be stopped," accompanied by the rhetorical question: "How many poor people in the world must die before we consider this a 'crime against humanity'?"

(b) Around the same time, the US National Academies of Science, Engineering and Medicine (2017) pointed out that the International Rice Research Institute (IRRI) had stated reported: "Golden Rice will only be made available broadly to farmers and consumers if it is successfully developed into rice varieties suitable for Asia, approved by national regulators, and shown to improve vitamin A status in community conditions. If Golden Rice is found to be safe and efficacious, a sustainable delivery program will ensure that Golden Rice is acceptable and accessible to those most in need" (p. 228). As of July 2016, IRRI was continuing research on developing varieties of Golden Rice for use in SE Asia, and (according to it) none of the conditions it stated had yet been met - it is for this reason that Golden Rice has not been introduced.

(c) Two years later, earlier this year (2018), IRRI asked the USFDA for an opinion regarding the safety of a variety of Golden Rice (called GR2E - the only variety yet submitted for regulatory approval - but not yet approved in any Asian country). FDA (May 24, 2018) endorsed the evaluation of IRRI (and the Australian regulatory body) that GR2E is safe for consumption, while pointing out that it is not intended for food or animal uses in USA. However, it added: "the concentration Beta-carotene in GR2E rice is too low to warrant a nutrient content claim." GR2E is safe but not nutritionally relevant.

(d) The signers of the letter, thus, were remarkably uninformed about the state of research on Golden Rice – and also about the views and stances of Greenpeace (I am not associated with Greenpeace). On its website Greenpeace states that its objective is to "ensure the ability of Earth to nurture life in all its diversity." It fits into the body of critics of using GMOs, who maintain that the dominant food-agricultural system (in which using GEOs has become for the time being a fundamental component) cannot respond adequately to the food and nutrition needs of the world's impoverished peoples (and the right to food security for everyone), and that these needs can best be ameliorated by the programs of agroecology and food sovereignty (Lacey, 2015a; 2015b) – and that programs for developing GEOs (like Golden Rice) are taking resources away from developing effective and lasting solutions to death and disease caused by vitamin A deficiency. Greenpeace has a respected place among these critics (and its "direct actions" and contributions to legal challenges are often appreciated by them). Of course, it would be legitimate to rebut the critics with argument and evidence. One wonders why the laureates did not attempt to do so.

(e) The credibility of pronouncements made by scientists of outstanding achievement is weakened when they sign letters like this one, accompanied by inflated, emotionally charged rhetoric, that has has a slender basis in fact. It would be enhanced if they entered into the type of dialogue, advocated by Helen Longino, in which scientists would "listen to" the evidence provided by relevant parties, attempt to understand critics, and not tar them without a hearing. Science has an indispensable contribution to make in policy deliberations; but it is not the determiner of policy. Science will be enhanced, and its role in democratic societies consolidated, if it claims only to have authority where it is actually warranted.

Notes

i **DRAFT** (not for citation outside of the PSA meeting in Seattle) – October 15, 2018. The text is a draft of the presentation I'm planning to make. The notes contain details that will be incorporated into an eventual completed paper.

ii See Appendix.

iii E.g., Mark Lynas (Cornell Alliance for Science), *A plea to Greenpeace*, <http://www.marklynas.org/2016/06/a-plea-to-greenpeace/>.
In this paper I only consider GEOs used in agriculture. I take for granted that claims to the effect that using GEOs is safe refer to GEOs that have passed safety tests, including those currently available on the market. (Obviously an unsafe GEO could be developed. Some varieties of GEOs have been developed that, after failing to pass safety tests, were not released for use.)

iv More fully developed and defended in Lacey (2005), Part 2.

v Deliberations concerning (2) and (3) cannot be settled in scientific inquiry (sound empirical inquiry), but there are sound empirically-based inputs that are (or could be) relevant to them. The deliberations will not be satisfactory if they do not draw upon these inputs. (See Lacey, 2005.)

vi Claims about efficacy need to be stated in a more qualified and nuanced way. I also will not contest that the claim that scientific research has not provided compelling evidence that consuming GEO products is unsafe health-wise. (The absence of compelling evidence that GEO products are unsafe to consume does not mean that there is compelling evidence that they are safe to consume – it depends on whether or not the necessary research has been conducted.)

vii The *ceteris paribus* qualification is needed to take into account that sometimes considerations, not reducible to safety ones, may properly be appealed to.

viii I will not discuss here how this set is generated – e.g., from considering past investigations, role of values in it, stakeholders' concerns, etc) – and who (holding what values?) makes (and should make) the identification of what should be considered harmful? following what kinds of deliberations? and who should be represented in the deliberations?.

ix To conclude on the basis of empirical investigation that 'X is safe' requires showing one-by-one that each member of the set of anticipated effect (judged to be harmful) is unlikely to occur at sufficient magnitude under the conditions imposed by proposed regulations. This presupposes: (a) an inductive move to unanticipated effects; and (b) that representative cases of all the effects, that should be labelled potentially harmful, are members of the set.

x I have argued elsewhere that here methodological and value considerations mutually reinforce each other (Lacey, 2017). Proponents of using GEOs often say that these safety studies investigate the risks occasioned by the GEOs themselves, and not those occasioned by the accompaniments of using them in agroecosystems or by socioeconomic mechanisms.

xi E.g., effects on soil microorganisms, a matter especially salient for those who regard maintaining soil fertility as indispensable for sustainable agriculture.

xii The studies, which have produced many of the results that have actually informed public policy and regulatory decisions, have been criticized for having a number of kinds of shortcomings (e.g., connected with conflicts of interest, and the use of intellectual property rights to maintain studies secret and so unavailable for replication and independent confirmation). Value judgments pervade these criticisms and their rebuttals. I will not attend to the questions that arise here.

xiii Such challenges might be deemed irrelevant by those who reject the value-outlook for which the possible effects have special salience, and so who reject the need for the further studies. Those adhering to the values of capital and the market sometimes take such a stand. How reasonable that might be depends on the arguments offered against holding the value-outlook in question.

xiv For elaboration see Lacey (2016; 2017).

xv For documentation, see, e.g., Bombardi (2017); Paganelli, et al. (2010); Human Rights Watch (2018).

xvi After a jury in California recently ruled that Monsanto was responsible for a man's being afflicted with cancer, and imposed a huge fine on it because it – for it was deemed that Monsanto had "acted with malice" in not providing warning on its label of the risks to health occasioned by using Roundup – the President of Bayer (that has now incorporated Monsanto) responded: "The correct use of Roundup doesn't present a risk to health" (reference to be added). [Monsanto has appealed the ruling.]

xvii Three years ago, when representatives of farmers – who had been poisoned in this way – came to present their testimony at a meeting of the "technical" commission in Brazil (CTNBio) that had appraised a particular variety of GEOs as safe, they were not granted a hearing since (most members of the commission maintained) they were bearers only of anecdotal (not scientific) evidence that had no relevance to the conclusions of scientific safety studies. When they then disrupted the meeting (and others of their group prevented the planting of a new variety of GEOs by invading a nursery and pulling up all the seedlings), they were denounced by major scientific organizations as having no respect for science, and acting on the

Seattle, WA; 1-4 November 2018 -397-

basis of "emotion and dogma." For criticisms of this stance taken by the majority of members of CTNBio, and a response to a rebuttal of the criticism, see Lacey, et al. (2015a; 2015b), articles published in *JC Notícias*, a daily e-newsletter of *Jornal da Ciência*, a publication of SBPC (Brazilian Society for the Advancement of Science).

The narrow scope of "scientific safety studies" is sometimes justified on the ground that the investigations of the social impact of using GEOs is not "scientific," for the methodologies adopted in them are not reducible to those adopted in the mainstream areas of science mentioned above. Be that as it may: I won't quibble about how to use the term "scientific" (a thick ethical term); the investigations in question are (when properly conducted) systematic empirical investigations. If they don't count as "scientific", that would imply that the results of "scientific" investigations cannot provide sufficient input into deliberations concerning public policies about safety, and would need to be supplemented with input from other kinds of empirical investigations.

xviii See Lacey (2017).

xix For details, see Lacey (2005; 2015a; 2015b).

xx See, e.g., Hilbeck, et al. (2015); Krimsky (2015); Traavik & Ching (2007).

xxi See Appendix.

xxii Unless all the relevant research has been conducted (and it has not been in this case), the absence of compelling evidence that GEO products are unsafe to consume does not imply that there is compelling evidence that they are safe to consume – and it has nothing to do with harms that may be caused by, e.g., contact with an agrotoxic, rather than by consumption.

The Reference Class Problem for Credit Valuation in Science

Carole J. Lee (c3@uw.edu)

Abstract: Scholars belong to multiple communities of credit simultaneously. When these communities disagree about how much credit to assign to a scholarly achievement, this raises a puzzle for decision theory models of credit-seeking in science. The reference class problem for credit valuation in science is the problem of determining to which of an agent's communities – which reference class – credit determinations should be indexed for any given act under any given state of nature. I will identify strategies and desiderata for resolving ambiguity in credit valuation due to this problem and explain how pursuing its solution could, ironically, lead to its dissolution.

1. Introduction

Within the scientific community, there is a common understanding that its reward system drives problematic behavior linked to publication patterns, pipeline retention, hypercompetitive scientific cultures, and reproducibility. Conversely, there is also a shared sentiment that, in order to change these cultures and behaviors in ways that would improve science, the scientific community must coordinate across institutions to change how credit is assigned at the level of the individual scientist (Alberts et al. 2014, Nosek et al. 2015, Aalbersberg et al. 2017, National Academies of Sciences 2018, National Science Foundation 2015, Blank et al. 2017). The hope is

Seattle, WA; 1-4 November 2018 -399-

that increasing individual researchers' incentives towards increased transparency and openness will improve the integrity, reproducibility, and accuracy of the published record.[1]

Analogously, philosophers working in the "credit economy" tradition adopt the working assumption that there is some amount of credit that agents can accrue for different acts under different states of nature. This assumption allows them to use decision theory to model how credit-seeking among individual scientists can give rise to behavior and norms that support or thwart the achievement of community-wide goals. When, in the aggregate, individual credit-seeking cuts against collective ends, their approach can explore how changes to individuals' incentive structures can nudge and redirect individual behavior (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Bright 2017, Heesen 2017, Kitcher 1990, Strevens 2003, Zollman 2018). Different philosophers make different assumptions about the norms by which credit gets allotted – for example, whether credit is best thought of as all-or-nothing (Strevens 2003, Bright 2017, Heesen 2017) or as something that may come in degrees (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Zollman 2018). However, the general approach assumes that there is some precise way to assign credit to different acts under different states of nature – an assumption that allows these philosophers to model credit-seeking behavior and the emergence of scientific norms in formally tractable ways.

But, how much credit gets assigned to any given act under any given state of nature? Just as each of us simultaneously belongs to multiple social categories each of which is tied to implied social hierarchies (Macrae, Bodenhausen, and Milne 1995, Crenshaw 1989), each

[1] Institutions can also experience incentives that promote or thwart scientific ends (Lee and Moher 2017).

scholar simultaneously belongs to multiple communities of value with implied social hierarchies for assigning credit. To which of an agent's communities – which reference class – should credit determinations be indexed and why?

In this paper, I will use examples from the current context of science's complex and dynamic culture to motivate and illuminate what I will call the *reference class problem for credit valuation in science*. I will identify a few strategies and desiderata for solving ambiguity in credit assignments due to the reference class problem. And, I will say a bit about how developing the resources needed to solve it could ultimately sow the seeds for its own dissolution.

2. The Reference Class Problem for Credit Valuation in Science

The contours of this puzzle about the "coin of recognition" (Merton 1968, 56) become visible when one moves beyond thinking about credit in generic, abstractions of scientific communities towards the heterogeneous communities we find today. I start from this slightly more concrete perspective because prestige requires recognition *by individuals and forums* that are themselves valued by credit-seeking scholars (Zuckerman and Merton 1971, Lee 2013): credit worthiness in science is a function of the individuals and systems designed to assess, allocate, dispute, and enforce it. Although some aspects of Zuckerman and Merton's narrative about the origins of the normative structure of science have been contested by historians (Csiszar 2015, Biagioli 2002), we see the social dynamics Zuckerman and Merton proposed clearly at play in contemporary science. For example, Nature Publishing Group recently found that – for the 18,354 authors in science, engineering, and medicine surveyed – the reputation of a journal is the primary factor driving choices about where to submit their work, where reputation is

primarily determined by the journal's impact factor and whether it is "seen as the place to publish the best research" (Nature Publishing Group 2015). Factors associated with a journal's ability to archive and disseminate research – things like a journal's time from acceptance to publication, indexing services, or Open Access options – were much less important.[2]

Within academia, each of us simultaneously belongs to multiple communities of value. The reference class problem arises when these different communities of value disagree about the amount of credit an agent accrues for choosing some act under some state of nature. Although I take this problem is be general, for the sake of clarity and simplicity in presentation, I will focus my examples on communities that can be described as having a nesting structure: for example, individual scholars belong to specific sub-disciplines, which are nested within disciplines, which are nested within a more general population of scholars. A sub-population that is nested within a population can have a credit sub-culture whose valuations differ from that of the population, whose valuations can differ from that of the super-population. In these cases, changing how narrowly or broadly one draws the boundaries of an agent's community of valuation can change the amount of credit assigned to a scholarly accomplishment. This gives rise to the *reference class problem for credit valuation in science*: to which of the agent's communities – which reference class – should credit valuations be indexed when determining the amount of credit the agent accrues for different acts under different states of nature?

[2] I recognize that some decision theorists, especially those working outside of philosophy, may reject or remain agnostic about attributing mental states such as beliefs to agents (Okasha 2016). However, because I understand credit and credit-seeking as sociological phenomena involving status beliefs such as these, I am committed to attributing beliefs to agents.

There are many examples across academia where nesting community structures can give rise to paradoxes and pathologies in credit assignments. For example, scholars' individual sense of what counts as quality work – their individual credit assignments – may deviate from what is endorsed in a sub-discipline or discipline's status hierarchy (Correll et al. 2017, Centola, Willer, and Macy 2005, Willer, Kuwabara, and Macy 2009). A puzzle that has cachet in a sub-discipline may be of peripheral importance within that discipline: for example, a more accurate technique for measuring how temperature cools with elevation considered critical in mountain meteorology and mountain ecology (Mindner, Mote, and Lundquist 2010) may have less visibility, despite its relevance, to the larger discipline of hydrology (Livneh et al. 2013). A question or technique that is thought to have high impact across fields (e.g., machine learning) may have little prominence within some of those fields.

Hypothetically speaking, one could imagine differences in valuations giving rise to a *Simpson's paradox in credit valuation*. Simpson's paradox is a phenomenon whereby a trend that appears in a population reverses or disappears when it is disaggregated into sub-populations (Blyth 1972). For example, a classic study found that, when looking at aggregate graduate school admissions data at UC Berkeley, women were, on the whole, less likely than men to be accepted; however, when the data was disaggregated into admitting departments, women were more likely than men to be admitted (Bickel, Hammel, and O'Connell 1975). Analogously, a *Simpson's paradox in credit valuation in science* would occur in cases where a population-level preference for scholarly product *a* versus *b* reverses when the population is disaggregated into its component sub-populations. In Simpson's Paradox cases, thinking more carefully about the context of evaluation usually leads to using a reference class that is finer-grained than the population-level. However, it's not clear whether this would always be the case in evaluations of

Seattle, WA; 1-4 November 2018 -403-

scientific credit. Hypothetically speaking, consider a hypothetical scenario in which an interdisciplinary project is not preferred by the individual disciplines represented by its authors or content, but is preferred when those disciplines are aggregated together. And, imagine that this project gets published in a journal, valued by those disciplines, that seeks papers of interest *across and beyond disciplines* (not just within disciplines): this is one way to interpret, for example, *Science's* mission to publish papers that "merit recognition by the wider scientific community and general public. . . beyond that provided by specialty journals" (Science). Which reference class would be most relevant in evaluating the value of this project?

There are other ways of dividing scholarly communities into nesting structures that create tensions in credit assignments. The pressures a scholar may feel from the incentive structure impacting her department/school may be slightly different from the incentive structure impacting her university. A coarse but concrete way to see this is to think about the prestige structure reified and reinforced by ranking systems (Espeland and Sauder 2012, 2016, Sauder and Espeland 2006), which transform "the ways professional opportunities are distributed" (Espeland and Sauder 2016, 7). An untenured business school professor with a potentially high impact manuscript needs to burnish her prestige in the eyes of both her dean and her provost, since both will evaluate her tenure case. If her provost is working to gain stature on the Academic Rankings of World Universities [ARWU], the professor should submit her manuscript to *Science* or *Nature*, since the ARWU ranks universities by their publications in these journals (Academic Ranking of World Universities 2018). However, if her dean is trying to gain stature on the *Financial Times* International ranking of MBA programs, she should submit to one of the fifty business, economics, or psychology journals by which the FT ranking system evaluates Business

school prestige – notably, the journal list does not include *Science* or *Nature* (Ormans 2016). What should the business school professor do?

Finally, credit assignments can vary depending on how long a time window a scholar keeps in view. A coarse but concrete way to think about this is by looking at how metrics for evaluating scholarship change over time. Journal impact factors are becoming less useful measures for evaluating an individual's scholarly contribution: since the advent of the digital age, the most elite journals (including *Science* and *Nature*) are publishing a decreasing percentage of the top cited papers (Larivière, Lozano, and Gingras 2013); the relationship between journal impact factor and paper citations has declined over time (Lozano, Larivière, and Gingras 2012); and, the citation distributions between journals "overlap extensively" (Larivière et al. 2016). The current wisdom is that if quantitative indicators are to be used to evaluate research, it is more useful to use article-level metrics such as citations as well as alternative metrics such as downloads and views (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017). On the horizon, there are now calls for creating new metrics that can encourage researchers and journals to be transparent and open in their reporting practices (National Academies of Sciences 2018, Wilsdon et al. 2017, Aalbersberg et al. 2017). Note that, the rise of such metrics – as well as the growing meta-research literature that ranks journals by the replicability (Schimmack 2015) or sample size and statistical power of their published results (Fraley and Vazire 2014) – makes it possible for a journal's impact factor and epistemic credibility to come apart (Fang and Casadevall 2011).

Decision theorists capture the risky nature of individual choices by allowing for uncertainty about which states of the world will come to be; and, when the probabilities attached to different outcomes are understood subjectively, these models permit a kind of subjectivity in

Seattle, WA; 1-4 November 2018 -405-

estimates of expected credit for different acts. However, I hope the examples throughout this section animate genuine *ambiguity in credit* due to the reference class problem for credit valuation in science.

3. Strategies and Desiderata for Solving the Reference Class Problem

How might decision theorists try to solve the reference class problem for assigning credit in science? One possible approach argues for the "correctness" of using one community rather than another. For example, it might be tempting to argue that all prestige is discipline-based since many scholarly prizes are distributed for excellence in particular disciplines (e.g., Nobel prize, Fields prize, academic society prizes); and, even when research is funded or published in interdisciplinary contexts, it may be primarily evaluated on the basis of its disciplinary excellence (Lamont 2009, but see Lee et al. 2013). Indexing credit valuation to a particular community need not prevent scholars from outside that community from understanding the relative value of that contribution: for example, if one were to adopt the old-fashioned and problematic assumption that an article's impact can be measured by the impact factor of the journal in which it is published,[3] and one recognizes that citations rates vary across disciplines, one could use field-normalized percentiles to understand a paper's impact in a metric that is legible across fields (Hicks and Wouters 2015). Because this strategy for addressing the

[3] The citation distributions within journals are so skewed that it is statistically improper to infer the impact of an individual article on the basis of the impact factor of the journal in which it is published (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017, Larivière et al. 2016, Wilsdon et al. 2015).

reference class problem relies heavily on identifying the "right" community, defending the centrality of the chosen community as opposed to others is critical. For example, some may challenge the idea that disciplines should be the sole arbiter of credit: note that the awarding of some scientific prizes reach across disciplinary conceptions of excellence (e.g., consider winners of the MacArthur Genius Prize and the psychologists who have won the Nobel Prize in Economics).

Another possible approach creates an algorithm that calculates the credit value of a scholarly contribution by summing the credit valuation of multiple communities. This approach would need to identify exactly how much to weight each community's valuation – with a rationale for why – since different weightings could lead to different overall credit valuations.[4] Note that some scholars take this style of approach when trying to measure the relative prestige of journals: in particular, the Eigenfactor score rates journals according to the number of its incoming citations, where the "relative importance" of each incoming citation is contextualized by the frequency with which the citing journal is itself cited (West, Bergstrom, and Bergstrom 2010).

Those who may wish to model the implications of different approaches for solving the reference class problem may try to do so by setting up hypothetical communities that assign

[4] On the face of it, this may seem like a form of commensuration because it involves summing values to calculate an overall score (Espeland and Stevens 1998). However, the process of commensuration requires combining values across *qualitatively* different domains of value. For clearer examples of commensuration in scholarly evaluation, see Lee (2015).

community boundaries and credit assignments in *de facto* ways to see what kinds of behaviors and norms emerge.

However, to solve the underlying conceptual problem, one must provide theories of community and credit that address two fundamental but vexing questions. How should one define and gerrymander the boundaries of the relevant communities invoked in the proposed solution? And, how does one determine the amount of credit those communities would assign to different acts under different states of nature? These questions may not be independently answerable. The boundaries of a community may need to be defined in terms of patterns of shared lore among its members about how credit is accrued – shared beliefs that coordinate credit-seeking and enforcement behavior in cases where status beliefs are internalized as norms (Merton 1973) and in cases where they are not (Willer, Kuwabara, and Macy 2009, Ridgeway and Correll 2006). Conversely, in recognition that some community members can have more influence than others on the content of reigning status beliefs, a community's credit assignments may need to be defined with some reference to the causal patterns of interaction among specific individuals and clusters of individuals – including status judges who wield "social control through their evaluation of role-performance and their allocation of rewards for that performance" (Zuckerman and Merton 1971, 66). Note, however, that answers to these questions should not *exclusively* inform each other. Notably, we must be careful not allow the size of a scholarly population and/or the power of its status judges to fully determine the intellectual value of the questions pursued by any particular partition of the scholarly universe.

4. Conclusion

Scientific credit – the "coin of recognition" (Merton 1968, 56) – is assessed, allocated, disputed, and enforced by many different communities and institutions within science that support and sustain a multiplicity of status hierarchies. This gives rise to what I have called the reference class problem for credit valuation in science. Solving this problem requires developing rich theories of community and credit that are based on fine-grained information about the structure and status systems of complex scholarly networks. The irony of this assessment is that such investigation towards solving the reference class problem could ultimately sow the seeds for its own dissolution.

In particular, such study can render friable a critical assumption for both the reference class problem and for decision theory models: namely, that communities, once defined, assign determinate amounts of monistic credit for different acts under different states of nature – that credit "can vary quantitatively but not qualitatively" (Anderson 1993, xii).[5] Contrary to this, recent policy papers call for moving away from narrowly conceived measurements of research excellence towards broader ones that are sensitive to the diversity of individual researchers', programs', and academic institutions' research missions (Hicks and Wouters 2015, Wilsdon et al. 2015). Such work can include community-engaged scholarship that creates, disseminates, and implements knowledge in coordination with the public to identify social interventions, change social practice, and influence policy (Hicks and Wouters 2015, San Francisco Declaration on Research Assessment 2013, Boyer 1990, Escrigas et al. 2014). From the

[5] Note too that, for formal reasons, the assumption that individual credit assessments could be aggregated into a collective one is questionable given the challenges of combining individual preferences into collective ones (Arrow 1950).

Seattle, WA; 1-4 November 2018 -409-

perspective of these efforts, plurality in our notions of scholarly excellence and credit – and differences in valuation and prioritization practices between individuals and communities – may be best conceived, not as a logical problem to solve, but as a starting point for theorizing.

Acknowledgments: Many thanks to Christopher Adolph, Aileen Fyfe, Crystal Hall, Jessica Lundquist, Conor Mayo-Wilson, and Kevin Zollman for helpful conversations. This research used statistical consulting resources provided by the Center for Statistics and the Social Sciences, University of Washington.

References

Aalbersberg, IJsbrand Jan, Tom Appleyard, Sarah Brookhart, Todd Carpenter, Michael Clarke, Stephen Curry, Josh Dahl, Alex DeHaven, Eric Eich, Maryrose Franko, Len Freedman, Chris Graf, Sean Grant, Brooks Hanson, Heather Joseph, Véronique Kiermer, Bianca Kramer, Alan Kraut, Roshan Kumar Karn, Carole Lee, Aki MacFarlane, Maryann Martone, Evan Mayo-Wilson, Marcia McNutt, Meredith McPhail, David Mellor, David Moher, Alison Mudditt Mudditt, Brian Nosek, Belinda Orland, Tim Parker, Mark Parsons, Mark Patterson, Solange Santos, Carolyn Shore, Dan Simons, Bobbie Spellman, Jeff Spies, Matt Spitzer, Victoria Stodden, Sowmya Swaminathan, Deborah Sweet, Anne Tsui, and Simine Vazire. 2017. "Making science transparent by default; Introducing the TOP Statement." *OSF Preprints*. doi: https://doi.org/10.31219/osf.io/sm78t.

Academic Ranking of World Universities. 2018. "ShanghaiRanking's Academic Ranking of World Universities 2018 Press Release." accessed September 1.

http://www.shanghairanking.com/Academic-Ranking-of-World-Universities-2018-Press-Release.html.

Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. 2014. "Rescuing US biomedical research from its systematic flaws." *Proceedings of the National Academy of Sciences* 111 (16):5773-7.

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Arrow, Kenneth J. 1950. "A difficulty in the concept of social welfare." *Journal of Political Economy* 58 (4):328-46.

Biagioli, Mario. 2002. "From Book Censorship to Academic Peer Review." *Emergences: Journal for the Study of Media & Composite Cultures* 12 (1):11-45.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. "Sex bias in graduate admissions: Data from Berkeley." *Science* 187 (4175):398-404.

Blank, Rebecca, Ronald J. Daniels, Gary Gilliland, Amy Gutmann, Samuel Hawgood, Freeman A. Hrabowski, Martha E. Pollack, Vincent Price, L. Rafael Reif, and Mark S. Schlissel. 2017. "A new data effort to inform career choices in biomedicine." *Science* 358 (6369):1388-9.

Blyth, Colin R. 1972. "On Simpson's Paradox and the sure-thing principle." *Journal of the American Statistical Association* 67 (338):364-66.

Boyer, Ernest L. 1990. *Scholarship Reconsidered*. San Francisco, CA: The Carnegie Foundation for the Advancement of Teaching.

Bright, Liam Kofi. 2017. "On Fraud." *Philosophical Studies* 174:291-310.

Bruner, Justin, and Cailin O'Connor. 2017. "Power, Bargaining, and Collaboration." In *Scientific Collaboration and Collective Knowledge*, edited by Thomas Boyer-Kassem, Conor Mayo-Wilson and Michael Weisberg, 135-157. Oxford, UK: Oxford University Press.

Centola, Damon, Robb Willer, and Michael Macy. 2005. "The emporer's dilemma: A computational model of self-enforcing norms." *American Journal of Sociology* 110 (4):1009-40.

Correll, Shelley J., Cecilia L. Ridgeway, Ezra W. Zuckerman, Sharon Jank, Sara Jordan-Bloch, and Sandra Nakagawa. 2017. "It's the conventional thought that counts: How third-order inference produces status advantage." *American Sociological Review* 82 (2):297-327.

Crenshaw, Kimberle. 1989. "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics." *University of Chicago Legal Forum* 139:139-168.

Csiszar, Alex. 2015. "Objectivities in Print." In *Objectivity in Science: New Perspectives from Science and Technology Studies*, edited by Flavia Padovani, Alan Richardson and Jonathan Y. Tsou, 145-69. Cham, Switzerland: Springer International Publishing.

Escrigas, Cristina, Jesús Granados Sánchez, Budd Hall, and Rajesh Tandon. 2014. "Editor's introduction. Knowledge, engagement and higher education: Contributing to social change." In *Report: Higher Education in the World*, edited by Cristina Escrigas, Jesús Granados Sánchez, Budd Hall and Rajesh Tandon. Palgrave Macmillan.

Espeland, Wendy Nelson, and Michael Sauder. 2012. "The Dynamism of Indicators." In *Governance by Indicators: Global Power through Quantification and Rankings*, edited by Kevin Davis, Angelina Fisher, Benedict Kingsbury and Sally Engle Merry, 86-109. Oxford: Oxford University Press.

Espeland, Wendy Nelson, and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York, NY: Russell Sage Foundation.

Espeland, Wendy Nelson, and Mitchell L. Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24:313-43.

Fang, Ferric C., and Arturo Casadevall. 2011. "Retracted Science and the Retraction Index." *Infection and Immunity* 79 (10):3855-9.

Fraley, R. Chris, and Simine Vazire. 2014. "The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power." *PLOS ONE* 9 (10):e109019. doi: 10.1371/journal.pone.0109019.

Heesen, Remco. 2017. "Communism and the Incentive to Share in Science." *Philosophy of Science* 84:698-716.

Hicks, Diana, and Paul Wouters. 2015. "The Leiden manifesto for research metrics." *Nature* 520:429-31.

Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* LXXXVII (1):5-22.

Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Larivière, Vincent, Véronique Kiermar, Catriona J. MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. 2016. "A simple proposal for the publication of journal citation distributions." *BioRxiv*:062109.

Larivière, Vincent, George A. Lozano, and Yves Gingras. 2013. "Are elite journals declining?" *Journal of the Association for Information Science and Technology* 65 (4):649-55.

Lee, Carole J. 2013. "The limited effectiveness of prestige as an intervention on the health of medical journal publications." *Episteme* 10 (4):387-402.

Lee, Carole J. 2015. "Commensuration bias in peer review." *Philosophy of Science* 82:1272-83.

Lee, Carole J., and David Moher. 2017. "Promote Scientific Integrity via Journal Peer Review." *Science* 357 (6348):256-7.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in peer review." *Journal of the American Society for Information Science and Technology* 64 (1):2-17.

Livneh, Ben, Eric A. Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M. Andreadis, Edwin P. Maurer, and Dennis P. Lettenmaier. 2013. "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions." *Journal of Climate* 26 (23):9384-9392.

Lozano, George A., Vincent Larivière, and Yves Gingras. 2012. "The weakening relationship between the Impact Factor and papers' citations in the digital age." *Journal of the American Society for Information Science and Technology* 63 (11):2140-45.

Macrae, C. Neil, Galen V. Bodenhausen, and Alan B. Milne. 1995. "The Dissection of Selection in Person Perception: Inhibitory Processes in Social Stereotyping." *Journal of Personality and Social Psychology* 69 (3):397-407.

Merton, Robert K. 1968. "The matthew effect in science." *Science* 1968:56-63.

Merton, Robert K. 1973. "The normative structure of science." In *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Norman W. Storer, 267-78. Chicago, IL: University of Chicago Press.

Mindner, Justin R., Philip W. Mote, and Jessica D. Lundquist. 2010. "Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains." *Journal of Geophysical Research: Atmospheres* 115. doi: https://doi.org/10.1029/2009JD013493.

National Academies of Sciences, Engineering, and Medicine,. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, D.C.: The National Academies Press.

National Science Foundation. 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. In *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*.

Nature Publishing Group. 2015. "Author Insights 2015 Survey."

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Mahlhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. "Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility." *Science* 348 (6242):1422-5. doi: 10.1126/science.aab2374.

Okasha, Samir. 2016. "On the interpretation of decision theory." *Economics & Philosophy* 32 (3):409-33.

Ormans, Laurent. 2016. "50 Journals used in FT research." accessed September 1. https://www.ft.com/content/3405a512-5cbb-11e1-8f1f-00144feabdc0.

Ridgeway, Cecilia L., and Shelley J. Correll. 2006. "Consensus and the creation and status beliefs." *Social Forces* 85 (1):431-53.

Rubin, Hannah, and Cailin O'Connor. 2018. "Discrimination and Collaboration in Science." *Philosophy of Science* 85:380-402.

San Francisco Declaration on Research Assessment. 2013. "The San Francisco Declaration on Research Assessment (DORA)." accessed September 1. https://sfdora.org/read/.

Sauder, Michael, and Wendy Nelson Espeland. 2006. "Strength in numbers? The advantages of multiple rankings." *Indiana Law Journal* 81 (1):205-27.

Schimmack, Ulrich. 2015. "Replicability Ranking of 26 Psychology Journals." January 18. https://replicationindex.wordpress.com/2015/08/13/replicability-ranking-of-26-psychology-journals/.

Science. "Mission and Scope." accessed September 1. http://sciencemag.org/about/mission-and-scope.

Strevens, Michael. 2003. "The role of the priority rule in science." *Journal of Philosophy* 100 (2):55-79.

West, Jevin D., Theodore C. Bergstrom, and Carl T. Bergstrom. 2010. "The Eigenfactor Metrics™: A network approach to assessing scholarly journals." *College & Research Libraries* 71 (3):236-44.

Willer, Robb, Ko Kuwabara, and Michael W. Macy. 2009. "The False Enforcement of Unpopular Norms." *American Journal of Sociology* 115 (2):451-90.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. 2015. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.

Wilsdon, James, Judit Bar-Ilan, Robert Frodeman, Elisabeth Lex, Isabella Peters, and Paul Wouters. 2017. Next-generation metrics: Responsible metrics and evaluation for open science. Report of the European Commission Expert Group on Altmetrics. European Commission.

Zollman, Kevin J. S. 2018. "The Credit Economy and the Economic Rationality of Science." *The Journal of Philosophy* 115:5-33.

Zuckerman, Harriet, and Robert K. Merton. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System." *Minerva* 9 (1):66-100.

Pragmatism and the content of quantum mechanics

Peter J. Lewis

Draft – please don't quote

Abstract

Pragmatism about quantum mechanics provides an attractive approach to the question of what quantum mechanics says. However, the conclusions reached by pragmatists concerning the content of quantum mechanics cannot be squared with the way that physicists use quantum mechanics to describe physical systems. In particular, attention to actual use results in ascribing content to claims about physical systems over a much wider range of contexts than countenanced by recent pragmatists. The resulting account of the content of quantum mechanics is much closer to quantum logic, and threatens the pragmatist conclusion that quantum mechanics requires no supplementation.

1. Introduction

Quantum mechanics is, notoriously, a theory in need of interpretation. But there is very little agreement on what kind of interpretation it needs. That is, there is very little agreement concerning what the foundational problems of quantum mechanics *are*, and without such agreement, there is little hope for a consensus concerning what an acceptable solution to the problems might look like.

Here is a way to divide up the territory. We can distinguish between *descriptive* and *normative* questions concerning quantum mechanics. Descriptive questions concern what quantum mechanics *says*—the *content* of the theory, as expressed in textbooks and used in labs. Normative questions concern what quantum mechanics *should* say—and in particular, whether it should say something different from what it actually does say.

All parties to the debates over the foundations of quantum mechanics would agree, I think, that there is a legitimate descriptive question concerning the content of quantum mechanics. Even those philosophers and physicists who think that quantum mechanics wears its interpretation on its sleeve at least feel the need to correct the mistaken impressions of *other* philosophers and physicists concerning what quantum mechanics says. The normative question presupposes an answer to the descriptive one: some think quantum mechanics is just fine the way it is, others contend that it needs to be replaced or supplemented with something radically different, and in large part this difference in attitude depends on prior differences concerning the answer to the descriptive question.

As an illustration, consider a fairly standard narrative concerning the descriptive and normative questions. Descriptively speaking, quantum mechanics depends on a distinction between measurements and non-measurements: measurements follow one dynamical law, the collapse dynamics, and non-measurements follow a different dynamical law, the Schrödinger dynamics. Since these two dynamical processes are incompatible, a precise formulation of quantum mechanics requires a precise dividing line between measurements and non-measurements. Quantum mechanics nowhere provides such a thing—and indeed, it seems highly unlikely that a term like “measurement” could be given a physically precise definition. So

Seattle, WA; 1-4 November 2018 -419-

descriptively speaking, quantum mechanics is inadequate as a physical theory. On the basis of this measurement problem, Bell (2004, 213–231) recommends replacing quantum mechanics with either a pilot-wave theory or a spontaneous collapse theory. For similar reasons, Wallace (2012, 35) recommends replacing quantum mechanics with a many-worlds theory.[1]

But not everybody concurs. There are alternative narratives according to which quantum mechanics, descriptively speaking, is just fine as it is, and hence there is no normative pressure to supplement or replace it. One prominent version proceeds from the quantum logic of von Neumann (1936) and Putnam (1975) through to the quantum information theory of Bub (2016). According to this approach, quantum mechanics describes a non-classical event space—in terms of truth values, a non-Boolean algebra, and in terms of probability ascriptions, a non-simplex distribution. No-go theorems (arguably) show that it is impossible to construct a set of events obeying classical Boolean logic or classical Kolmogorov probability that reproduces the empirical predictions of quantum mechanics. The implication is that in quantum mechanics we have discovered something important about the fundamental event structure of the world. Seeking to replace or supplement quantum mechanics with a theory obeying classical logic and classical probability theory amounts to a quixotic attempt to impose a structure on the world that it manifestly does not have (Bub 2016, 222). The measurement problem, on this account, results from a mistaken demand for a dynamical explanation of the individual events in the quantum structure, when no such explanation is available (Bub 2016, 223)

[1] Wallace takes the many-worlds theory to be a precise statement of the content of quantum mechanics, rather than a replacement for it. I take up the question of whether the many-worlds structure is present in quantum mechanics as it stands in section 2.

This fundamental difference of opinion—between those who take the measurement problem seriously and those who regard it as a pseudo-problem—continues to divide the foundations of physics community today. Hence the descriptive question—the question of what quantum mechanics actually *says*—remains a pressing one. In this paper, I argue for a particular way of approaching the descriptive question. The methodology is the pragmatist one of Healey (2012; 2017) and Friederich (2015), but the answer to the descriptive question that results from following this methodology, I argue, differs in an important way from the answers that Healey and Friederich give. I conclude by assessing the consequences of this answer to the descriptive question for the normative question.

2. The descriptive question

So how should we approach the descriptive question? Consider a straightforward realist approach to the content of scientific theories. A theory, at least in physics, is typically expressed using a particular mathematical structure. The *state* of a physical system is generally identified with a mathematical entity that resides in a particular abstract space, and the *dynamics* of the theory tell us how that state evolves over time. So, for example, in many applications of classical mechanics, the state of a physical system can be represented by a set of vectors in a three-dimensional Euclidean space, and the dynamical laws of Newtonian mechanics tell us how the set of vectors evolves over time. The interpretation of the mathematics is fairly straightforward: the vectors represent the positions and momenta of point-like particles, and classical mechanics tells us how the properties of the particles change.

Seattle, WA; 1-4 November 2018 -421-

Such an approach can equally be applied to quantum mechanics (Albert 1996). According to quantum mechanics, the state of a physical system is identified with a complex-valued function defined on a configuration space—a space with three dimensions for each particle in the system. A dynamical law, the Schrödinger equation, tells us how this function, the wave-function, changes over time. Then by analogy with classical mechanics, the wave-function must be a representation of the physical properties of the quantum system as they change over time.

The continuity with classical mechanics in the above account is attractive, but there are surprising consequences. For an *N*-particle system, the wave-function is defined over a 3*N*-dimensional configuration space, and it cannot be represented without loss in a three-dimensional space. This has led some to conclude that a straightforward realist reading of quantum mechanics shows that the three-dimensionality of our physical world is illusory (Albert 1996). Furthermore, if we model a measurement using quantum mechanics, the wave-function ends up with components corresponding to each possible outcome of the measurement—not just one outcome, as is the case classically. This leads Everettians like Wallace (2012) to conclude that a straightforward realist reading of quantum mechanics shows that every possible outcome of a measurement actually occurs.

These conclusions might be right, but do they simply follow from close attention to the structure of quantum mechanics? There are reasons to be suspicious. As Healey (2017, 116) notes, conclusions of this kind depend on the assumption that the wave-function plays the same descriptive role in quantum mechanics as the position-momentum vectors play in classical mechanics. If this assumption is itself up for grabs in the interpretation of quantum

mechanics, then neither of these conclusions is warranted. But how do we adjudicate the question of whether the wave-function describes physical systems or whether it has some other, non-descriptive role? Is there a metaphysically neutral methodology that could be used to answer this question? Healey (2012; 2017) and Friederich (2015) think that there is.

3. Pragmatism

Consider an analogy. "Stealing is bad" has the same grammatical structure as "Cherries are red". But it is far from clear that both sentences should be taken as descriptive. In particular, badness, taken as a property of actions, seems like a queer kind of property, imperceptible and disconnected from the other properties of the action. Expressivists seek to dissolve the problem of the nature of badness by claiming that a sentence like "Stealing is bad" should be taken as expressive rather than descriptive—as expressing our attitude towards stealing. Pragmatists further coopt expressivism as a variety of pragmatism (Price 2011, 9). Pragmatists stress the variety of uses of language, noting that sentences with superficially similar form can be used in radically different ways. "Cherries are red" is used to describe a class of objects, whereas "Stealing is bad" is used to express our attitude towards a class of actions.

Pragmatism, then, enjoins us to pay close attention to how a sentence is *used* in order to find out what it means. Healey (2012; 2017) and Friederich (2015) each suggest that the pragmatist approach provides us with a metaphysically neutral methodology for probing the content of quantum mechanics. That is, we can look at how various quantum mechanical claims are used by physicists in order to determine what those claims mean. This strikes me as a welcome suggestion. In the rest of this section I present the conclusions of their pragmatist

Seattle, WA; 1-4 November 2018 -423-

inquiries; in the next, I consider whether the language use of physicists actually supports those conclusions.

Healey (2012) distinguishes between *quantum claims* and *non-quantum magnitude claims*. The former explicitly mention quantum states, quantum probabilities, or other novel elements of the theory of quantum mechanics. The latter are claims about the magnitude of a physical quantity that do *not* involve quantum states, quantum probabilities etc. In keeping with the pragmatist methodology, Healey bases this distinction on the way the two kinds of claims are used. Non-quantum magnitude claims are used in a straightforwardly descriptive way. But quantum claims are used in a different way: they are used, not to *describe* a system, but to *prescribe* a user's degrees of belief in various non-quantum magnitude claims.

As an example, Healey appeals to the Interference experiments of Juffmann et al. (2009), in which C_{60} molecules are passed through an array of slits and then deposited on a silicon surface. To derive quantum mechanical predictions for this experimental arrangement, quantum states are ascribed to C_{60} molecules. That is, quantum claims of the form "The molecule has state $|\psi\rangle$" are used, via the Born rule, to ascribe probabilities to claims concerning the various possible locations of the molecules on the silicon surface. These latter claims—of the form "The molecule is located in region R"—are non-quantum magnitude claims. The job of the non-quantum magnitude claims is to describe the physical system, but the job of the quantum claims is to prescribe degrees of belief in the non-quantum magnitude claims for an appropriately situated observer. In this respect Healey's approach is like the expressivist's in ethics: claims that have superficially similar grammatical forms have very different functions.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -424-

Another important strand in the pragmatist approach concerns the role of decoherence. After the C_{60} molecule hits the silicon surface, complicated interactions with the surface mean that the state of the molecule-environment system becomes approximately diagonal when written as a density matrix in the position basis. This in turn insures that the probabilities ascribed by the Born rule to various claims about the molecule's position closely obey the probability axioms. But before the molecule encounters the silicon surface, its state is a coherent superposition—a state that is not even approximately diagonal, and for which the Born rule does not ascribe probabilities to location claims that closely obey the probability axioms. For such a state, the Born rule does not prescribe appropriate degrees of belief in the non-quantum location claims, and so assertion of such claims prior to decoherence is not *licensed* by quantum mechanics. Decoherence, then provides a demarcation between situations in which it is appropriate to have a well-defined degree of belief in a non-quantum magnitude claim, and situations in which it is not.

The central finding of the Healey-Friederich pragmatist approach is that attention to the use of quantum mechanical language shows that claims about the quantum state of a system are not used to describe that system. Hence, we should not think of the wave-function as a representation of the physical properties of the quantum system as they change over time. This perspective has the advantage that the measurement problem does not arise: if the wave-function doesn't represent the system, then we don't have to worry that the dynamical laws for wave-function evolution are different for measurements and non-measurements. In fact, if the quantum state is prescriptive, then the difference between measurements and non-

Seattle, WA; 1-4 November 2018 -425-

measurements arises quite naturally: the results of measurements have a direct and obvious influence on what you should believe.

Hence the pragmatist approach provides a clear answer to the descriptive question: quantum mechanics, in itself, says *nothing* about the world. As Healey (2017, 12) puts it, "quantum theory has no physical ontology". Rather, quantum mechanics tells us what to believe about non-quantum ontology—about particles, or in the case of quantum field theory, about fields. Furthermore, this answer to the descriptive question suggests an answer to the normative question: since the measurement problem doesn't arise, there is no motivation for supplementing or replacing quantum mechanics with something else.

4. Actual use, counterfactual content

Thus far, I have said little about the evidence that backs up Healey's claims about how quantum claims and non-quantum magnitude claims are used. Indeed, direct evidence from the language use of physicists is likely to be unenlightening: that a claim is asserted in a given context provides no direct evidence concerning whether its content is descriptive or prescriptive.

To fill this gap, Healey appeals to an inferentialist account of the link between use and meaning derived from the work of Robert Brandom (2000): the meaning of a claim is identified with the set of material inferences it licenses. So by looking at the way a claim is used in licensing inferences, we can gain evidence about what it means. And here the distinction between prescriptive quantum claims and descriptive non-quantum magnitude claims seems to be well motivated. In the practice of physics, a claim about the quantum state of a system is

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -426-

used to infer Born probabilities, and nothing more. If Born probabilities are taken to be rational degrees of belief, then the prescriptive content of a quantum claim exhausts its meaning.

A non-quantum magnitude claim, on the other hand, can license a wide variety of inferences. From the claim that a C_{60} molecule is located in a particular region of the silicon surface, we can infer that an electron microscope will produce an image of the molecule if directed at that region (Juffmann et al. 2009, 2). We can infer that if the silicon surface is left untouched for two weeks, the C_{60} molecule will remain in the same place (Juffmann et al. 2009, 2). Under suitable conditions, we can infer that the C_{60} molecule will emits photons; under different conditions, that it will act as a nucleation core for molecular growth (Juffmann et al. 2009, 3). In other words, the inferences licensed by the non-quantum magnitude claim support the interpretation that the meaning of the claim is descriptive rather than merely prescriptive.[2]

So there is a good case to be made, I think, that actual use supports the distinction between prescriptive quantum claims and descriptive non-quantum magnitude claims. But there is a further strand to the Healey-Friederich interpretation, namely that non-quantum magnitude claims are only licensed after decoherence. This claim, I think, does not stand up so well to scrutiny.

Consider C_{60} interference again. After the molecule has adhered to the silicon surface, the state of the molecule is decoherent, and the claim that the molecule has a particular

[2] There is a sense in which the meaning of *any* claim is prescriptive according to the inferentialist program: the claim about the location of the molecule licenses an inference to a certain *degree of belief* that the electron microscope will produce an image of it. But still, there is a reasonable distinction here: the quantum claim licenses inferences only via the Born rule, whereas the non-quantum magnitude claim licenses inferences via a huge variety of schema typical of small physical objects. The latter is just what it is for a claim to be descriptive.

Seattle, WA; 1-4 November 2018 -427-

location is licensed—that is, it is appropriate to associate a particular degree of belief with the claim, and if that degree of belief is high enough, it is appropriate to assert the claim. But before the molecule has adhered to the silicon surface, the state of the molecule is coherent, and no claim about the location of the molecule is licensed—it is not appropriate to associate a degree of belief with such a claim, or to assert it. Similar considerations apply to properties other than location.

This seems to fly in the face of actual use. For example, in the description of the C_{60} interference experiment, Juffmann et al. (2009, 2) assert that "all transmitted particles arrive with the same speed," and "about 110cm behind the source, the molecules encounter the first diffraction grating," apparently ascribing both speed and location to C_{60} molecules prior to decoherence. This doesn't seem to be an isolated incident: physicists routinely talk of preparing, selecting, spraying, shooting and trapping particles, ions and molecules, and this talk typically involves making claims about these objects prior to any eventual decoherence.

It is possible, of course, that this is just "loose talk", or an indirect way of making claims about the quantum state of the systems concerned. But given the frequency of such claims, and given the reliance of the pragmatist methodology on *use*, this seems like a shaky game to play. It would be better, all things considered, if such claims could be accommodated within the pragmatist interpretation, rather than explained away as anomalies.

But there are obvious barriers to licensing non-quantum magnitude claims prior to decoherence. As Friederich (2015, 79) notes, the Born rule is only "reliable" when applied to decoherent states, in the sense that only for such states are the numbers it produces guaranteed to closely obey the probability axioms. Given some reasonable assumptions about

rationality, it is plausible that numbers that do not closely obey the probability axioms could not be rational degrees of belief. Furthermore, Healey argues that asserting a non-quantum magnitude claim prior to decoherence is likely to be misleading. For example, suppose one asserts (with Juffmann et al.) that "about 110cm behind the source, the molecules encounter the first diffraction grating." One might infer from this that each molecule passes through exactly one slit in the grating, and hence that the presence of the other slits is irrelevant, and hence that there is no possibility of interference (Healey 2012, 745).

So the pragmatist approach seems to face a dilemma: either it fails to accommodate the actual language use of physicists, or it licenses misleading assertions and irrational degrees of belief. Isn't there another way? I think there is. Consider a mundane claim like "There is beer in the fridge." In typical contexts, an assertion of this claim licenses the inference that if you were to go to the fridge and open the door, you could take a beer and drink it. Of course, you might not actually do this; maybe you don't want a beer. That is, the inference here is a counterfactual one. A good deal of the inferential content of our assertions has this counterfactual character.

Now return to the quantum context. Consider again the claim that "about 110cm behind the source, the molecules encounter the first diffraction grating." What content could that claim have? If we broaden the notion of inferential content to include counterfactual inferences, then the content seems fairly clear: if we were to replace the first diffraction grating with a detector taking up the same region of space, then the Born rule would ascribe a degree of belief close to 1 to detecting the molecules.

Seattle, WA; 1-4 November 2018 -429-

How does the inclusion of counterfactual content avoid the barriers to licensing non-quantum magnitude claims prior to decoherence? Note that the counterfactual content of the claim about the molecules involves a counterfactual intervention on the system—a counterfactual measurement. The counterfactual measurement induces counterfactual decoherence. The Born probabilities are conditional on this intervention and the associated decoherence, so the Born probabilities for various position claims concerning the molecules are not, after all, unreliable, in the sense of violating the probability axioms.

Neither should there be any danger of being misled by an assertion that the C_{60} molecules encounter the grating, because the counterfactual conditions implicit in the content of that assertion are distinct from the conditions that actually obtain in the apparatus. That you *could* detect the molecules at the diffraction grating, given a different experimental arrangement, doesn't license the inference that there *is* no interference, given the actual experimental arrangement. Admittedly, though, this amounts to a weakening of the content of position claims from the classical case, as spelled out in the next section.

5. A happy convergence?

I have argued that non-quantum magnitude claims have assertible content in a far wider range of contexts than countenanced by Healey or Friederich. If there is some counterfactual intervention on a system that would produce decoherence in the basis defined by a given observable, then claims about the values of that observable have content. And since counterfactual interventions only have to be realizable in principle, this means that claims about the value of an observable for a system *generally* have content, whether or not the

system *actually* decoheres in the basis defined by that observable. This has the welcome consequence that the frequent assertions made by physicists about the properties of systems prior to decoherence are contentful.

A potential cost of such permissiveness about content is that the structure of this content is, in general, non-Boolean. Consider again a C_{60} molecule that is approaching the first diffraction grating, and consider an assertion of "The molecule passes through the leftmost slit". This assertion has content, on the proposed view, because in principle there is an intervention on the system that would produce decoherence in a basis defined by an observable that distinguishes which slit the molecule passes through. Still, assertion of the claim would not be appropriate, simply because there are many slits in the grating, so the Born rule ascribes it a low probability. The same goes for every other slit in the grating. Nevertheless, the assertion that "The molecule passes through the leftmost slit, or the second to the left, or…" is assertible, since the Born rule ascribes it a probability close to 1. The disjunction is assertible, but none of the disjuncts is assertible. Since assertibility is a surrogate for truth in the pragmatist context, this is equivalent to saying that the disjunction is true, but none of the disjuncts is true.

One might take this to be unacceptable on the pragmatist view—especially if you endorse an inferentialist pragmatism, as Healey does. From a disjunctive claim you can straightforwardly infer that at least one of the disjuncts is true. If the content of a claim is identified with the inferences that it licenses, then part of the meaning of the disjunctive claim about the C_{60} molecule is that some assertion of the form "The molecule went through slit *x*" is true. Hence my proposal about content threatens to violate the inferentialist account of

Seattle, WA; 1-4 November 2018 -431-

meaning. The pragmatist interpretation of Healey and Friederich avoids this problem by insisting that claims about systems have meaning only after suitable decoherence.

Of course, pragmatism is not necessarily tied to an inferentialist account of meaning. But even given inferentialism, there is arguably no real problem here. Physicists are *selective* in the inferences they draw: from the disjunctive claim, they don't infer that the C_{60} molecule goes through some particular slit, so they don't infer a lack of interference. But they do infer that the molecule will arrive at the silicon surface, that it might radiate a photon in flight, and so forth. That is, the inferences drawn by physicists from their claims about pre-decoherent systems suggest that the non-Boolean structure of those claims is already *built in* to the meanings associated with those claims and revealed in inference.

This suggests that close attention to the way non-quantum magnitude claims are actually used leads to a happy convergence between pragmatism and the quantum logical approach. Physicists assert claims about particles even when the state does not decohere, and such claims seem to be meaningful. But physicists are not inclined on that basis to draw all the inferences that a full Boolean structure to their claims would license. Quantum mechanics apparently weakens the meaning of many claims about pre-decoherent physical systems, but without rendering those claims meaningless.

6. The normative question

As a methodology for addressing the *descriptive* question of the content of quantum mechanics, the pragmatist approach seems entirely appropriate: look to the *use* of physicists to determine what the various claims involved in the theory mean. At the hands of Healey and

Friederich, this approach yields the important insight that while non-quantum magnitude claims are used to describe physical system, quantum claims are used to prescribe appropriate degrees of belief in non-quantum magnitude claims. But Healey and Friederich go further, in limiting the assertibility of non-quantum magnitude claims to contexts in which the quantum state is decoherent in the relevant basis. This, I have argued, cannot be squared with the actual use of such claims. I propose instead that non-quantum magnitude claims *generally* have well-defined content, understood in terms of a counterfactual intervention on the system. This change to the pragmatist approach means that it ends up looking a lot like the quantum logical approach that preceded it. Indeed, the pragmatist approach might be regarded as a *justification* for quantum logical claims concerning the content of quantum mechanics.

But where does all this leave the *normative* question concerning whether quantum mechanics is fine as it is, or whether it should be supplemented or replaced? Healey and Friederich argue that quantum mechanics is fine as it is: if quantum claims do not describe physical systems, then there can be no conflict between the way that quantum mechanics describes systems during measurements and the way it describes them during non-measurements. If there is no measurement problem, then there is no motivation to replace such a successful theory. If, as Healey (2017, 12) maintains, quantum theory "states no facts about physical objects or events," then there can be no requirement that we come up with an *explanation* of quantum facts and events.

However, I have suggested that quantum theory has more content than the pragmatists countenance. In one sense, I agree that quantum theory states no facts: a quantum claim, such as the attribution of a quantum state to a system, is not a description. But in another sense,

Seattle, WA; 1-4 November 2018 -433-

there are distinctive quantum facts, or at least facts with a distinctive quantum structure: non-quantum magnitude claims about pre-decoherent systems exhibit the non-Boolean structure characteristic of quantum mechanics. This is the sense in which quantum logic gets things right.

Notably, though, the proponents of quantum logic *also* often take the view that quantum logic dissolves the measurement problem (e.g. Putnam 1975, 186). But this dissolution is widely regarded to be a failure (e.g. Bacciagaluppi 2009, 65) Once one has admitted that the structure of true (i.e. assertible) claims for a quantum system is non-Boolean, the question of *how* the world manages to instantiate this structure becomes legitimate and pressing. A denial that any explanation is required looks suspiciously like instrumentalism. And since any answer to this question goes beyond quantum mechanics as it stands, the call for explanation involves a demand to supplement quantum mechanics, or to replace it with something more fundamental.

Of course, given the no-go theorems, the path to an explanation of the structure of quantum facts is by no means clear. But neither do the no-go theorems show that an explanation is *impossible* (Friederich 2015, 161).[3] If the foregoing is correct, then pragmatism is an excellent way to *expose* the foundational problems of quantum mechanics, but it is not a means to *dissolve* them.

References

[3] Interestingly, Friederich (2015, 161) suggests supplementing quantum mechanics with sharp values for all observables, even though this seems at odds with his therapeutic aim of dissolving the foundational problems of quantum mechanics rather than solving them (2015, 6).

Albert, David Z. (1996), "Elementary quantum metaphysics," in J. T. Cushing, A. Fine and S. Goldstein (eds.), *Bohmian Mechanics and Quantum Theory: An Appraisal*. Dordrecht: Springer, 277-284.

Bacciagaluppi, Guido (2009), "Is logic empirical?" in K. Engesser, D. M. Gabbay and D. Lehmann (eds.), *Handbook of Quantum Logic and Quantum Structures.* Amsterdam: North-Holland, 49-78.

Bell, J. S. (2004), *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press.

Brandom, R. (2000), *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Bub, Jeffrey (2016), *Bananaworld: Quantum Mechanics for Primates*. Oxford: Oxford University Press.

Friederich, Simon (2015), *Interpreting Quantum Theory: A Therapeutic Approach*. Basingstoke: Palgrave Macmillan.

Healey, Richard (2012), "Quantum theory: a pragmatist approach," *British Journal for the Philosophy of Science* 63: 729-771.

Healey, Richard (2017), *The Quantum Revolution in Philosophy*. Oxford: Oxford University Press.

Juffmann, T., Truppe, S., Geyer, P., Major, A. G., Deachapunya, S., Ulbricht, H., and Arndt, M. (2009), "Wave and particle in molecular interference lithography," *Physical Review Letters* 103: 263601.

Price, Huw (2011), *Naturalism Without Mirrors*. Oxford: Oxford University Press.

Putnam, Hilary (1975), "The logic of quantum mechanics," in *Mathematics, Matter and Method: Philosophical Papers Volume 1*. Cambridge: Cambridge University Press.

von Neumann, John (1932), *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer-Verlag.

Wallace, David (2012), *The Emergent Multiverse*. Oxford: Oxford University Press.

Tool Migration: A Framework for Analyzing Cross-disciplinary Use of Mathematical Constructs

Chia-Hua Lin
University of South Carolina / KLI

Abstract. Mathematical formalisms that are constructed for inquiry in one disciplinary context are sometimes applied to another, a phenomenon that I call 'tool migration.' Philosophers of science have addressed the advantages of using migrated tools. In this paper, I argue that tool migration can be epistemically risky. I then develop an analytic framework for better understanding the risks that are implicit in tool migration. My approach shows that viewing mathematical constructs as tools while also acknowledging their representational features allows for a balanced understanding of knowledge production that are aided by the research tools migrated across disciplinary boundaries.

Keywords: Cross-disciplinarity, tool migration, epistemic risks

1. Introduction

Mathematical formalisms that are constructed for scientific inquiry in one disciplinary (or sub-disciplinary) context are applied to another. Philosophers of science have started paying attention to this cross-disciplinary aspect of scientific practice. For instance, the discussion of 'model transfer' concerns a relatively small set of mathematical models that are applied in multiple disciplinary contexts. Humphreys (2004) proposes that models that are transferred to study phenomena of a different domain owe their versatility to the computational tractability they afford. In contrast, Knuuttila and Loettger (2014, 2016) suggest that in addition to tractability, versatile models also offer conceptual frameworks for theorization, which they label 'model templates.' However, these analyses do not deal with the risks inherent in this aspect of scientific practice. Consider the use and development of game theory in evolutionary biology as an example. In importing game theory, which was originally conceived to describe strategic interaction between rational agents typically studied by social scientists, evolutionary biologists may need to modify the theory in order to generate knowledge about presumably non-rational agents, at least in many cases. One can then assume that any changes to the theory--between its established applications in social sciences and its novel uses in evolutionary biology--require special attention so as to avoid misinterpreting an analysis.

Despite the advantages, there might be risks associated with using mathematical constructs across disciplines. In this paper, I ask: might there be patterns of transfer that may undermine the effectiveness of the imported mathematical formulation? What would these

patterns, if any, look like? This paper is an attempt to explore the conditions in which importing mathematical constructs may be epistemically risky. To begin, I develop a framework to systematically characterize the landscape of mathematical importations. The goal of such a framework is two-fold. Proximately, the framework captures characteristics of migration that the current terminology, such as 'model transfer' or 'importing/exporting,' fails to discern. Ultimately, with this additional discernibility, I suggest that one may start to explore and identify patterns of importation that may be subject to epistemic risks, such as misinterpretation of an outcome produced by using a imported mathematical construct.

In Section 2, I argue that one can view mathematical constructs in science in terms of 'research tools' and that transporting such tools across disciplines, which I call 'tool migration,' can in some cases be a disservice to science. Next, I classify tool migration based on two kinds of contextual details that bear significance to the effectiveness of the migrated research tool in a foreign context. In Section 3, I apply this approach to the use and development of game theory in evolutionary biology. Finally, in Section 4, I discuss in what ways this tool migration framework, which is essentially a typology of four types of tool migration, may help to characterize epistemically risky patterns of tool migration.

2. Theoretical Background

Although the notion of epistemic risks associated with migration of mathematical constructs has not been explicitly addressed, the idea of viewing mathematical constructs as research tools follows from the discussion on the ontology of scientific models. Ever since the shift of attention to scientific practice (e.g., Hacking 1983), there has been a growing literature in which models in science are viewed as entities *detachable* from theory and data (e.g., Morrison 1999; Morgan and Morrison 1999). One recent predecessor to my tool migration account is a pragmatic approach to scientific models put forth by Boon and Knuuttila (2008). In their paper, which uses examples from engineering, they argue that scientific models are better understood as 'epistemic tools' instead of as representations of some target systems in the world. Boon and Knuuttila's argument draws heavily on the epistemological roles of scientific models in relation to the scientists who use them. According to them, scientific models allow their users "to understand, predict, or optimize the behavior of devices or the properties of diverse materials" (2008, 687). Thus, for an ontological account of scientific models to be productive and realistic, as they argue, it should be sensitive to the relation between the models and the modelers, i.e., the tools and their users. An adequate evaluation of Boon and Knuuttila's argument will take us far afield, but my work will show that both the representational and the pragmatic aspects are indispensable to a better understanding of the epistemic risks in tool migration.

2.1 Viewing mathematical constructs as research tools

In general terms, any mathematical construct that is to be *used or operated* in an algorithmic manner, and the outcome of whose operation is to be *interpreted* in order to answer a research question, is an example of what I am calling a research tool. Let me first unpack the operational aspect of a research tool.

Let's assume that the proper use of any mathematical constructs employed in scientific research is expected to produce consistent results. To achieve this consistency, then, a well-defined procedure needs to accompany such a construct so that anyone who follows the procedure expects, and is expected, to obtain the same outcome given the same input. For instance, when performing a game-theoretic analysis, one goes through a sequence of steps, such as: (i) identify the players and the acts available to them, (ii) identify the payouts in every set of acts, (iii) find the 'Nash equilibria,' which refers to a set of acts, one for each player, in which no player could improve his or her payoff by unilaterally changing act. A similar algorithmic procedure can be seen when applying, say, Newton's law of gravitation:

$$F_{grav} = G\frac{m_1 m_2}{r^2}. \qquad (1.1)$$

For example, the sequence of steps to obtain the magnitude of the gravitational force, F_{grav}, between any two objects includes: (i) identify the mass of each object, (ii) identify the distance between them, (iii) complete the equation in which 'm_1' and 'm_2' refer to the masses of the two objects, 'r' the distance in between, and 'G' the gravitational constant. In these two examples, when the first two steps produce consistent input, the third step is expected to generate the same output.

Moreover, concerning the interpretational aspect of a research tool, the output of a series of symbol assignments and manipulations can be understood *only through the lens of some interpretation*. The Nash-equilibrium of a game is a meaningful 'solution' in virtue of the usual understanding of the game-theoretic formulation of a problem. Similarly, the meaning of the value obtained through completing the equation in (1.1) is derived from the usual interpretation of the quantities appearing in the equation and the theoretical context in which those quantities are defined.

Finally, assume that something can be viewed as a tool if it serves as a means to an end. In this case, then, mathematical constructs like game theory or mathematical formulas can be seen as research tools. In the case of applying a mathematical construct, the goal of performing a sequence of prescribed steps goes beyond merely completing the calculation and obtaining a result. Instead, the output is to be interpreted so that one may solve a problem, answer a research question, or gain knowledge about a subject-matter. Thus, a mathematical construct that prescribes algorithmic symbol manipulation can be seen as a research tool, assisting its users to meet an end. Manipulating symbols is a means to the end that was specified during the mathematical formulation of the research problem.

Seattle, WA; 1-4 November 2018 -439-

2.2 Epistemic risks of tool migration

Another predecessor to my account is Morgan's discussion of the re-situating of knowledge (2014). According to her, knowledge production is necessarily 'situated,' and consequently, applying a piece of knowledge outside its initial context requires effort - different contextual situations require different 're-situating' strategies. The term 're-situation' thus captures what scientists do in practice to transport locally generated knowledge across contexts. As she argues, to make an instance of scientific knowledge accessible outside its production site, one needs to establish inferential links between the production site and the destination site. However, she suggests, whether a re-situation of knowledge contributes to scientific progress depends on whether the transport secures some sort of inferential safety.

Building from Morgan's notion of the re-situation of knowledge, I argue that cross-disciplinary use of research tools is epistemically risky. Given the locality of scientific knowledge production, applying scientific knowledge outside its production site may come with epistemic risks. For example, between the production site and a destination site, there may be incongruent disciplinary characteristics (e.g., implicit theoretical assumptions) that fail to be captured by the inferential strategy, such that knowledge from the former cannot be transferred to the latter. Similarly, we can assume that the construction of a research tool is also *situated* in nature. Namely, a research tool is conceived to be operated and to extend our knowledge concerning a subject-matter *given a particular disciplinary context*. It follows that cross-disciplinary use of research tools is as epistemically risky as re-situating knowledge. That is, the epistemic reliability (i.e., general ability or tendency to produce knowledge) of some research tool in one disciplinary context does not necessarily carry over to another.

The concept of 'tool migration' captures both the 'situated-ness' of a research tool that was established in its native discipline and the effort it takes to 're-situate' the tool in a foreign discipline. Naturally, in the process of uprooting a research tool, significant contextual details—ranging from implicit expertise to important background assumptions—may be stripped away. Likewise, during re-situation, new features may be introduced to the tool so as to treat a different subject matter in a new disciplinary context. Together, due to the possibility of losing or gaining significant contextual details, or both, a cross-disciplinary tool migration risks undermining the effectiveness of the tool. These risks include, for example, misinterpretation of the research result or failure to produce genuine knowledge. Thus, it follows that tool migration can in some cases be a disservice to the production of knowledge.

Acknowledging these challenges, some have argued against the cross-disciplinary effort to integrate disciplinary knowledge (e.g., van der Steen 1993). Alternatively, one might try to overcome these challenges so long as the risks are better understood and managed. To understand the risks, I suggest that we first look at the patterns of tool migration. Among these patterns, we might find that some of them could be epistemically risky. Having established the

notions of research tools and risks involved with tool migration, I turn to the contextual details that are closely related to a tool's epistemic performance.

2.3 Contextual details of a research tool: the target profile and the usage profile

The construction of a research tool is necessarily situated within a context. In order to compare and contrast between the native (or established) context and the foreign context of a migrated tool, I single out two major types of details.

The first type concerns the assumptions about the entities that are studied by a subject-matter for which the tool is developed. For instance, game theory defines what it considers as a game, a player, or an act. For simplicity, I call *all* the assumptions that a tool makes about its target entities the tool's 'target profile.'

The second type considers *the ways* in which one interprets the output from applying a tool in his or her research. In a game-theoretic analysis, for example, by following an algorithmic procedure, one obtains a solution of a game in the form of a Nash equilibrium. Depending on the game that one was analyzing, the solution could be understood as an explanation of economic behavior, or a prediction about it, or it could be used to optimize an strategic interaction. For simplicity, I call *all* the ways in which a tool is intended to be used, e.g., describing, predicting, optimizing, or explaining its target phenomenon, the tool's 'usage profile.'

Together, as I demonstrate in Section 4, the 'target profile' and 'usage profile' allow one to detect patterns of changes in the contextual details between the established use and the novel use of a research tool. They are able to do this because these two profiles offer a coarse resolution; looking through the lens of the target profile and usage profile, one zooms out from particular cases of tool migration so as to detect patterns of cross-disciplinary transport. Further analyses of these patterns will then shed lights on their associated epistemic risks.

2.4 Four types of tool migration

With the two profiles of a research tool and the two contexts in which the tool is used, i.e., a novel use and an established use, one can distinguish four types of tool migration.

First, compared to its established use, when a novel use of a tool catalyzes changes in both target and usage profiles, the tool migration is transformative, and therefore I call it a ***tool-transformation***. Second, in contrast, when both target and usage profiles remain more or less intact after the migration, the tool's novel use is considerably similar to its previous applications. Thus, I call such a case ***tool-application***. Between these two extreme types, there are novel uses of a research tool that alter only one of the two profiles but not both. When a tool changes its target profile but not its usage profile, I call it a ***tool-transfer,*** and when a tool changes its usage profile but not the target profile, I call it a ***tool-adaptation***. See **Table 1** for a summary.

Seattle, WA; 1-4 November 2018 -441-

Table 1
A Typology of Tool Migration

Between established and novel uses of a research tool	**Usage profile remains**	**Usage profile deviates**
Target profile remains	'Tool-application'	'Tool-adaptation'
Target profile deviates	'Tool-transfer'	'Tool-transformation'

Among these four types of tool migration, tool-transfer is arguably the most familiar to the philosophers of science. Humphreys coins the term ‘computational templates’ to refer to a relatively small number of mathematical equations that are applied to investigate different domains of phenomena (2002, 2004). Bailer-Jones (2009) discusses such a scientific practice in terms of mathematical analogy. For one example, Newton's law of gravitation was intentionally sought after to model electrostatic force (see Bailer-Jones 2009 for a detailed account). The important parallel between the two formulas, shown in (1.2), is that both types of forces (gravitational and the electrostatic) are proportional to the inverse of the square of the distance, r, between two masses, m_1 and m_2, or two charges, q_1 and q_2. The constants that appear in both formulas scale the quantities to match empirical phenomena.

$$F_{grav} = G\frac{m_1 m_2}{r^2} \quad \text{and} \quad F_{el} = k\frac{q_1 q_2}{r^2} \tag{1.2}$$

In contrast, the other three types of tool migration, despite prominent examples, are less explored in regard to their general features. One prominent example of tool-transformation is the development of game theory to be used in evolutionary biology.

3. The Migration of Game Theory From Social Sciences to Biology

In this section, I show in what sense the novel use of game theory in evolutionary biology, which is now known as 'evolutionary game theory' ('EGT') can be considered as a tool-transformation. I should mention that my account of the migration of game theory in this paper is not meant to address all the limitations of both game theory and EGT in their respective disciplinary contexts. Instead, the purpose of this account is to show that one *can* detect patterns of migration that have epistemic implications by focusing on the target profile and usage profile of a research tool.

3.1 Game theory in social sciences

Game theory was initially formulated to mathematically model strategic interactions between intelligent, rational agents. In game theory, a game is defined as an interaction between two or

more players in which each player's payoff (e.g., profit) is affected by the decisions made by other players. Typically, such a game assumes both *perfect information* and *common knowledge. Perfect information* assumes that all players know the entire structure of the game (all moves and all payouts) as well as all previous moves made by all players in the game (if it is an iterated or multi-move game). *Common knowledge* is the assumption that all players know that all players have perfect information, and that all players know that all players know that all players have perfect information, and so on. That is, *common* knowledge concerns what players know about what other players know. Moreover, the players also recognize that all players are cognizant that all players are rational, i.e., there is common knowledge of the game and of the *unbounded rationality* of all players. As such, all players will act in the way that takes all other players' potential moves into account in order to maximize their odds of winning. In addition to these assumptions regarding the players of a game, the structure of a game, which refers to the combinations of each move and its payout, is usually summarized in a 'payoff matrix.' Typically, an analysis of a game aims to find out its 'solution,' a unique Nash equilibrium (or sometimes equilibria) of the game.

Game theory has been used in economics, as well as other social sciences, to describe, predict, optimize, or explain a variety of human interactions, such as the economic behaviors of firms, markets, and consumers (e.g., Brandenburger and Nalebuff 1995; Casson 1994) military decisions (Haywood 1954) or international politics (e.g., Snidal 1985).

3.2 Game theory in evolutionary biology

Game theory was later used in evolutionary biology, where a game is understood as phenotypes (or heritable traits) in contest. In 1973, John Maynard Smith and George Price borrowed the formalism of a payoff matrix from game theory to mathematically model the evolution of phenotype frequencies in a population of organisms (see Grüne-Yanoff 2011). Their modeling method assumed that phenotypes are in contest with other phenotypes in a population of organisms. For instance, in a Hawk-Dove game, the contest is embodied by organisms with the phenotype of being aggressive and other organisms that are peaceful. In such a context, the payoff of a move is interpreted as the reproductive success of the phenotype (i.e., the number of copies it will leave to the next generation). Moreover, while the terminology such as 'game,' 'payoffs' and the formalism of a payoff matrix can be seen in the novel use of game theory in biology, the solution to a game in evolutionary biology is decidedly different from the Nash-equilibrium. An evolutionary game theoretic analysis typically looks for an evolutionarily stable strategy (ESS), i.e., a distribution of phenotypes in a population that is stable.

3.3 Epistemic implications of tool transformation

It is clear that the target profile of game theory is no longer the same between its established use

Seattle, WA; 1-4 November 2018 -443-

in social sciences and its novel use in biology. First, none of the assumptions of *perfect information*, *common knowledge*, and *unbounded rationality* in what is now known classical game theory (CGT) remain in the novel use of game theory in biology. Second, the moves in EGT are heritable phenotypes exhibited by a group of organisms instead of acts available to players. Third, the payoffs in EGT are the reproductive success of the heritable traits. In this sense, the three assumptions concerning the players were stripped away from the tool - as a result of uprooting game theory from social sciences, and the *heritability* assumption about the moves as well as Darwinian fitness interpretation of the payoff were introduced to the tool - as a result of re-situating it to evolutionary biology.

Note that the change in the target profile forces a limitation to the usage profile of the migrated tool. For instance, nullifying the *unbounded rationality* assumption concerning the players, EGT can no longer be used to optimize a game, i.e., discovering the rationally optimal strategy, which is a common use of game theory in social sciences. For instance, in the prisoner's dilemma, the Nash-equilibrium is for both players to defect. This solution is often interpreted as a prescription for the game; the players are irrational not to defect. However, in a Hawk-Dove game, the ESS obviously has no such normative use. Because the 'moves' of being an aggressive type or a peaceful type are not 'chosen,' the idea of there being normatively better or worse choice of moves is therefore questionable. Moreover, the organisms are not assumed to be rational. Thus, while the players in the prisoner's dilemma could be said to be irrational for choosing to cooperate, this sense of normativity does not carry over to the evolutionary game theoretic analysis of the Hawk-Dove game. One would be mistaken to say that it is 'irrational' for the doves to be doves. Thus, the change in the target-profile of game theory, especially the stripping away of the *unbounded rationality* assumption, has resulted in how the migrated tool should or should not be used.[1]

Moreover, applying EGT to study social phenomena (e.g., Axelrod 1984) or cultural evolution (e.g., Skyrms 2010) requires a careful re-defining of the terms (such as fitness) so as to avoid misinterpretation. Using EGT in social sciences, which can be considered as a 'homecoming' of the migrated tool, is not uncommon. However, the notion of payoffs in EGT refers to, roughly, the overall biological reproductive success of a group of organisms that exhibit a phenotype. Obviously in a social context, reproductive success of the members of some group is not, very often, the feature of interest. A careful reinterpretation of payoffs is thus needed in every analysis to prevent misleading conclusions.

[1] Of course, a more interesting prescriptive use of the ESS of a Hawk-Dove game might be, for example, to manage ecosystems for optimal predator-prey balance. Nevertheless, it should be noted that a justification for this type of prescriptive use of EGT would require further analysis because it is apparently not be derived from CGT.

To generalize, this example suggests that at least in some cases, a change in the target profile requires a corresponding change in the usage profile, or failure of producing genuine knowledge may follow. So far, I have shown that a solution of an ESS analysis may not be interpreted as an optimization to a Hawk-Dove game. Applying EGT to study social phenomena also requires careful treatment to the notion of payoff. Now if, hypothetically, some researcher were to make either of these two mistakes, his or her novel use of the tool would have been classified as tool-transfer - the novel use changes only the target profile without also changing the usage profile. It suggests that in some cases, tool-transformation may not be as risky as tool-transfer. I will come back to the issue of tool-transfer after some remarks related to the migration of game theory.

4. Contributions of the Tool Migration Analysis

The tool migration typology and its focus on tracking both similarities and differences meets the needs to sharpen discussions concerning inter- or cross-disciplinary use of research tools. Current literature seems to lack a framework to capture important, relational characteristics of the research tools that appear in multiple disciplinary contexts. For instance, 'tool-transformation' captures significant differences in details between CGT and EGT without losing sight of the contextual relationship between the two. In contrast, other terms in the literature, such as 'imports' or 'transfers,' fall short of doing so.

'Imports' signals the importation of research tools from a foreign discipline. In contrast, 'transfers' refers to the use of a scientific model, which was established to study phenomena of one domain, to study phenomena of a different domain. Neither term captures the migration of game theory to biology. As Grüne-Yanoff argues,

> [B]iologists constructed the more sophisticated formal [evolutionary game theoretic] concepts themselves. One could speak of the import of formal concepts only with respect to very basic notions such as strategies or pay-off matrices, and it may be more appropriate to refer to formal inspirations rather than imports or transfers in these contexts. (2011, 392)

Moreover, I have suggested that a change in a tool's target profile without a corresponding change in the tool's usage profile *may* lead to misinterpretation and hence misuse of the tool. If this observation is generalizable, which is debatable, then it follows that cases of tool-transfer are epistemically riskier than cases of tool-transformation. On the other hand, if this observation applies only to some cases, it nevertheless reveals at least two epistemic implications concerning tool migration: 1) when the target profile changes, one must be careful not to draw conclusions that might be natural in the old context but may not make sense within the new context, given the new target, and 2) sometimes a change in target profile can, force a change in usage profile. Potentially failing to recognize when these changes occurred in a migration leads

Seattle, WA; 1-4 November 2018 -445-

to risky uses of the migrated tool.

Morgan (2011) has argued that while not all scientific knowledge travels far, those that travel with integrity (i.e., maintaining their content more or less intact during its travels) and travel fruitfully (i.e., finding new users or new functions) are considered to be traveling well. It is relatively easy to quantify the latter feature – one needs to look at just the number of a tool's novel applications. However, determining whether a tool has traveled with integrity is not straightforward. As a starting point, this proposed tool migration framework—especially its distinction between the target profile and the usage profile of a tool—provides a starting point that is crucial for assessing the integrity of a migrated research tool. With this framework, one may discover more patterns of tool migration that impact the epistemic integrity and, consequently, effectiveness of a migrated research tool in a foreign discipline.

5. Conclusion

I have argued that mathematical constructs used in science can be viewed as research tools and their cross-disciplinary novel use as tool migration. I have also argued that making novel use of established tools has its risks, but such an implication is not meant to deter cross-disciplinary sharing of tools. Indeed, certain important breakthroughs in the history of science are due to creative, unconventional, uses of research tools (e.g., the use of Fourier's mathematical treatment of heat to study electrostatics [Thomson 1842] or the use of Faraday's mechanical model of fluid motion to model the electromagnetic field [Maxwell 1861]). Versatile research tools are not rare in science. A framework of tool migration aims to offer not only a useful terminology to characterize the diverse landscape of their versatility but also a groundwork to investigate risky patterns of making novel use of established research tools. Finally, this tool migration approach shows that viewing these constructs as tools whilst acknowledging their representational features (i.e., as captured in their target profile) allows for a balanced understanding of knowledge production - especially those productions that are aided by research tools that have migrated across disciplinary boundaries.

References

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Bailer-Jones, Daniela M. 2009. *Scientific Models in Philosophy of Science*. University of Pittsburgh Press.

Brandenburger, Adam M., and Barry J. Nalebuff. 1995. *The Right Game: Use Game Theory to Shape Strategy*. Harvard Business Review.

Boon, Mieke, and Tarja Knuuttila. 2009. "Models as Epistemic Tools In Engineering Sciences: A Pragmatic Approach." In *Handbook of the Philosophy of Science*, edited by Anthonie Meijers, 687–720. Elsevier B.V.

Casson, Mark. 1994. *The Economics of Business Culture: Game Theory, Transaction Costs, and Economic Performance*. Oxford University Press.

Grüne-Yanoff, Till. 2011. "Models as Products of Interdisciplinary Exchange: Evidence from Evolutionary Game Theory." *Studies in History and Philosophy of Science Part A* 42 (2): 386–97.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Haywood Jr, O. G. 1954. "Military Decision and Game Theory." *Journal of the Operations Research Society of America* 2 (4), 365–85.

Humphreys, Paul. 2002. "Computational Models." *Philosophy of Science* 69 (September): 1–27.

___ 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Knuuttila, Tarja, and Andrea Loettgers. 2014. "Magnets, Spins, and Neurons: The Dissemination of Model Templates across Disciplines." *The Monist* 97 (3). The Oxford University Press: 280–300.

Knuuttila, Tarja, and Andrea Loettgers. 2016. "Model Templates within and between Disciplines: From Magnets to Gases–and Socio-Economic Systems." *European Journal for Philosophy of Science* 6 (3). Springer: 377–400.

Maynard Smith, John, and George Price. 1973. "The Logic of Animal Conflict." *Nature* 246: 15–18.

Maxwell, James Clerk. 1861. "Xxv. on Physical Lines of Force: Part I.–the Theory of Molecular Vortices Applied to Magnetic Phenomena." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21 (139): 161–75.

Morgan, Mary, and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Vol. 52. Cambridge University Press.

Morgan, Mary. 2010. "Travelling Facts." In *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, edited by Peter Howlett and Mary Morgan, 3–39. Cambridge University Press.

___ 2014. "Resituating Knowledge: Generic Strategies and Case Studies." *Philosophy of Science* 81 (5). University of Chicago Press: 1012–24.

Morrison, Margaret. 1999. "Models as Autonomous Agents." In *Models as Mediators: Perspectives on Natural and Social Science*, edited by Mary Morgan and Margaret Morrison, 38–65. Cambridge University Press.

Skyrms, Brian. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press.

Snidal, Duncan. 1985. "The Game Theory of International Politics." *World Politics* 38 (1). Cambridge University Press: 25–57.

Thomson, William. 1842. "On the Uniform Motion of Heat in Homogeneous Solid Bodies and Its Connection with the Mathematical Theory of Electricity." *Cambridge Mathematical Journal* 3 (1842): 71–84.

Van Der Steen, Wim J. 1993. "Towards Disciplinary Disintegration in Biology." *Biology and Philosophy* 8 (3): 259–75.

Representations are Rate-Distortion Sweet Spots

Manolo Martínez (mail@manolomartinez.net)

Abstract

Information is widely perceived as essential to the study of communication and representation; still, theorists working on these topics often take themselves not to be centrally concerned with "Shannon information", as it is often put, but with some other, sometimes called "semantic" or "nonnatural", kind of information. This perception is wrong. Shannon's theory of information is the only one we need.

I intend to make good on this last assertion by canvassing a fully (Shannon) informational answer to the metasemantic question of what makes something a representation, for a certain important family of cases. This answer and the accompanying theory, which represents a significant departure from the broadly Dretskean philosophical mainstream, will show how a number of threads in the literature on naturalistic metasemantics, aimed at describing the purportedly non-informational ingredients in representation, actually belong in the same coherent, purely information-theoretic picture.

1 Information, Shannonian and Dretskean

In what follows I will use a random variable, S, to encode the state the world is in, and another random variable, M, for signals. How should we characterize the information that values of M (i.e., individual signals) carry about values of S (i.e., individual world states)? The most basic quantity with which information theory records dependence among two random variables is the *mutual information* between them. This quantity being an expected value, Dretske (1981, p. 52f) claims, renders it unsuitable for an analysis of representational status, and it should be substituted by notions that record relations between individual states, S_i, and individual signals, M_j. The basic relation which substitutes mutual information in contemporary Dretskean accounts is that of *making a probabilistic difference* (Scarantino 2015): a signal M_j makes a probabilistic difference to the instantiation of a state S_i iff the following *basic inequality* holds:

$$P(S_i|M_j) \neq P(S_i)$$

Nearly all the accounts of information developed in the recent, and not so recent, philosophical literature on this topic are variations on, and attempts to quantify, this inequality. For illustration, in Skyrms (2010, p. 36) the "information in $[M_j]$ in favor of $[S_i]$" is defined as the *pointwise mutual information* (Also *pmi* henceforth) between

state and signal. There is a direct relation between pmis and the basic inequality: the former are nonzero iff the latter is true.

The running thread connecting most prominent contemporary accounts of information is that all there is to Shannon's information theory, at least for the purposes of investigating the nature of representation, is two quantities: the unconditional probability of states and the probability of states conditional on signals, perhaps rearranged as the logarithm of their ratio, or in some other way. Unsurprisingly, from this it is routinely concluded that there is much more to representation than information. This conclusion is premature: informational content in the Dretskean tradition is not by a long shot all there is to information theory. This should not be taken to imply that information is all there is to representation—for one thing, I believe with teleosemanticists (Millikan 1984; Papineau 1987) that teleofunctions have a role to play in a complete theory of representation—but it does mean that no Dretske-style "semanticized information" needs to be recognized, over and above the quantities studied in information theory proper. I will argue that it also means that some prominent proposals as to ways to bridge the information-representation gap are, in fact, unwittingly appealing to informational structure.

In the following section I review two such proposals. My aim is not to argue against them—they are built upon largely correct insights. I will instead aim at showing that a better informed understanding of information provides a way to incorporating these insights in a unified, purely information-theoretic picture.

2 Bridging Information and Representation

2.1 Many-to-One-to-Many Architectures

The first proposal is that it is not enough that representations carry information; on top of that, they must sit in the right place in a certain cognitive architecture. Sterelny (2003), for example, has argued that the emergence of representations is enabled by two prior evolutionary transitions: from "detection" to "robust tracking", on the one hand; from "narrow-banded" to "broad-banded" behavioral responses, on the other. Robust tracking is in essence a *many-to-one* relation between world state and signal: many sensory inputs give rise to one and the same representation. Other theorists have advocated similar architectural constraints on representational vehicles. Famously, Burge (2010) places a great deal of weight on *perceptual constancies* in his characterization of perceptual representation (Burge 2010, p. 413.) This is a variation on Sterelny's idea and, as such, a many-to-one architectural constraint on representational status.

As for broad-banded responses, in these systems a single representation will be flexibly dealt with, resulting in different courses of action, depending on the context where the representation is tokened. Response breadth is in essence a *one-to-many* relation between representational vehicle and output: one representation, many agential outputs.

2.2 Reference Magnetism

A second proposal has been to focus on the entities that should figure in the content of simple representations. The suggestion, typically, is that represented entities should be appropriately *natural*, or *real*. For example, Dan Ryder (2004, 2006) has argued that neurons become attuned to *sources of correlation*. These entities are closely related to Richard Boyd's *homeostatic property clusters* (also HPC henceforth, Boyd 1989): HPC theory identifies natural kinds with clusters of properties which tend to be instantiated together, and such that this frequent co-instantiation is not just a statistical fluke. What Ryder calls sources of correlation are the grounds for these HPC-related frequent co-instantiations—whatever it is that makes them *not* statistical flukes. Ryder claims that many of the representations the brain trades in target sources of correlation. Martínez (2013) and Artiga (forthcoming) have made more general cases that simple representations preferably target HPCs (Martínez), or properties that best explain the co-occurrence of other properties (Artiga).

A similar idea has been explored in an entirely independent line of enquiry starting with Lewis (1983): "among the countless things and classes there are ... [o]nly an elite minority are carved at the joints, so that their boundaries are established by objetive sameness and difference in nature. Only these elite things and classes are eligible to serve as referents" (Lewis 1984, p. 227). This is what Sider (2014, p. 33) calls *reference magnetism*.

As I show in section 4, these two ideas, although apparently disparate, are in fact closely related, and the explanatory payback they bring to representation-involving talk depends on their informational underpinnings.

3 Information Theory is a Source-Channel Theory

Philosophy has understood information theory as a mostly *definitional* effort: for all philosophers have typically cared, the theory begins and ends with a presentation of what it takes for one random variable (or the worldly feature it models) to carry information about another. But information theory goes well beyond that. It is, well, a *theory*, and as such it is chiefly composed of claims that are advanced in the hope that they be true about the world.

In a nutshell, the most celebrated results in information theory have to do with specifying how faithful the transmission of information from a source can be, when it happens over a (typically noisy, typically narrow) channel. These results have played absolutely no role in informational accounts of representation.[1] Take, for starters, the idealized depiction of an information-processing pipeline in fig. 1 (*cf.* Cover & Thomas 2006, fig. 7.1)

[1] Two recent philosophical treatments of information that try to redress this neglect are Mann (2018) and Rathkopf (2017).

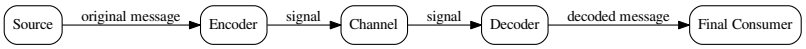


Figure 1: An information-processing pipeline

Here an *encoder* produces a signal as a response to information incoming from a source. This signal goes through a channel and is subsequently decoded, producing a message that is then utilized for whatever purposes downstream. The first thing to note is that the broadly Dretskean ideas about the content of a signal introduced in section 1 only have use for the first two links in this information-processing chain: how signals carry information about a certain original message produced by a source, as depicted in fig. 2. In fact, in information theory the main action happens immediately after that: a source is producing stuff, and we want that stuff to *go through a channel*. Information theory is mainly about providing theoretical guarantees of faithfulness in transmission, given the rate of the channel. We can think of this rate as the number of bits it provides for the encoder to use in the signal. If, say, the rate is 2 bits per use of the channel, this means the encoder can use up to 2 bits to construct the signal and be sure that it can pass unscathed through the channel and on to the decoder.

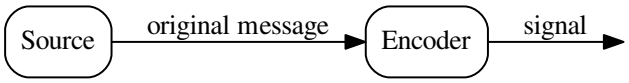


Figure 2: The information-processing pipeline in the Dretskean tradition

In typical cases of representation, channel rate is consistently smaller than ideal. Consider animal alarm calls. Vervet monkeys, for example, are typically described as being able to produce three different, discrete kinds of calls (Seyfarth, Cheney & Marler 1980a, 1980b) that are usually taken to be associated with the presence of leopards, eagles and snakes respectively. Obviously, the entropy of the relevant aspects of the environment that prompt the production of a call (think of all the possible patterns of approach of these predators, for example) vastly outstrip the rate of a channel, which consists in the production of just one out of three possible signals. This means that loss in communication is inevitable. Alarm calls, and for analogous reasons representations in general, are all about *lossy transmission*.

The way in which information theory deals with lossy transmission is by defining a *distortion measure* (Cover & Thomas 2006, p. 304) that gives a score to a pair composed of a certain original message M, and the decoded version thereof, $\hat{M}$. In what follows I

Seattle, WA; 1-4 November 2018 -451-

will be using the *Hamming distortion* which simply adds 1 to the distortion when the bits in the original and decoded signals (which we can assume to be binary strings) do not coincide, and 0 otherwise, then normalizes. So, for example, the Hamming distortion between an original signal $M = 010011$ and a decoded signal $\hat{M} = 100010$ is ⅓, because the first, second, and last (a total of 3) bits have been decoded incorrectly, and there are 6 bits in total.

The central result in this so-called *rate-distortion theory* approach to lossy transmission is that there is a *rate-distortion function*, $R(D)$, which gives the minimum rate at which any given distortion is achievable. The actual mathematical expression of the rate-distortion function need not detain us here (see Cover & Thomas 2006, p. 307, theorem 10.2.1), but it is such that the *Blahut-Arimoto* algorithm (Blahut 1972; Arimoto 1972) allows us to calculate it easily.

The main thesis of this paper is that representations belong in information-processing pipelines whose rate-distortion function has *sweet spots*: by this I mean points in the rate-distortion curve such that the usefulness of increasing the rate of the channel past those points is much smaller than before reaching them. Moreover, the encoding-decoding strategies that make use of these representations tend to live in the vicinity of those sweet spots. I submit that it is these information-theoretic properties that the conditions on representation discussed in section 2 try to get at.

To see how rate-distortion analyses work let's start by looking into a source that models a series of fair-coin tosses: this random variable would have two values, *heads* and *tails*, with associated probabilities $P_{heads} = P_{tails} = .5$). Using the Hamming distortion as our target distortion measure, if the coin lands heads (tails) and the decoded message is tails (heads) the distortion is 1, otherwise 0. The Blahut-Arimoto algorithm allows us to draw the rate-disortion curve, in fig. 3. Here the blue line is the rate-distortion curve. It intersects the x-axis at 1.0 bits (the entropy of the source) and it intersects the y-axis at 0.5 (the lowest average distortion one can achieve when the channel is closed.) The red line gives a measure of how steep the blue line is at any given point—in particular, the absolute value of the slope of the blue line. The higher the red line, the steeper the blue line.

The situation this setup is modeling is one in which a single cue is present or absent, and a signal tries to keep track of whether it does. This is precisely the kind of situation where many theorists (certainly Sterelny and Burge, for the reasons reviewed in 2.1) would see the postulation of representations as entirely idle—see, e.g., Schulte's vasopressin example in his Schulte (2015). In agreement with the idea that postulating representations here is idle, there is not much structure to the rate-distortion curve corresponding to this setup: reading the chart from right to left, increasing the rate makes the achievable expected distortion go smoothly down, until the rate hits the entropy of the source, at which point the achievable distortion is zero.That's about it.

Let's now model one kind of situation in which there is a reasonably wide consensus that representations make an explanatory contribution: vervet-monkey alarm calls, as reviewed above. In the model, the source—the situation the information-processing pipeline is dealing with—randomly makes members of two natural kinds (we can think

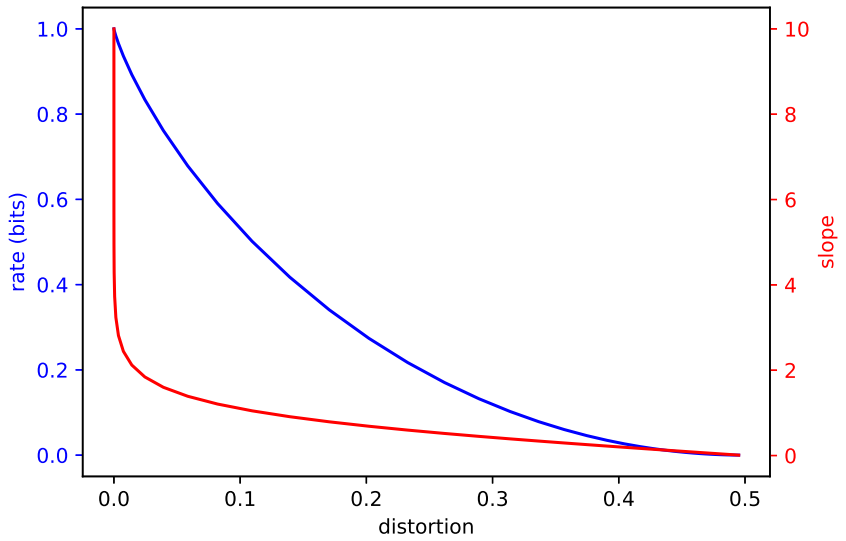


Figure 3: The rate-distortion function for a coin toss

of them as two different predators) be or not present at any given time, independently from one another. This intends to mimic the situation vervet monkeys face, where snakes, leopards and eagles show up or not, more or less at random.

These natural kinds are modeled as homeostatic property clusters (see section 2.2 above). In order to derive a explicit probability distribution for the source out of this qualitative description, the two HPCs are in their turn represented by two Bayesian networks, each with a parent node and four children (see fig. 4.) Each of the nodes stands for a property; if the node is *on* it means the corresponding property is instantiated; if it is *off* it means it is not. In the model, children nodes replicate noisily the state of their parent. Thus, e.g., if the parent is *on* (if the corresponding property is instantiated) each child property will have a .95 chance of being instantiated too; if the parent is *off* the probability for each children of being instantiated is .05. The unconditional probability of instantiation for the two parent nodes is .5.

In the model, the source produces a binary string, with each member of the string being 1 if the corresponding node is on, and 0 if it's off. This signal is encoded, goes through a channel, and is then decoded at the other side. The target distortion measure is the Hamming distortion. Fig. 5 plots the rate-distortion curve for this model.

This curve is very different from the one in fig. 3: there is a clear "sweet spot"—a sudden drop in the usefulness of extra rate, see the red curve—when the system hits a rate of 2 bit/use. I.e, there is, in a certain principled sense, an optimal level of lossy compression; a way to set up an encoding-decoding strategy that recover most of what's going on in

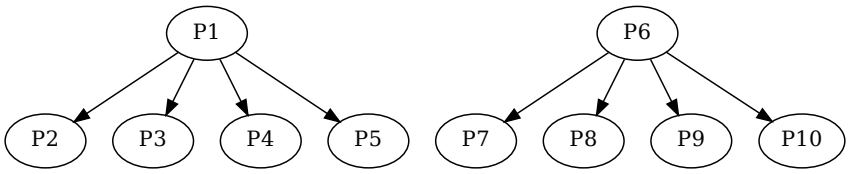


Figure 4: Two natural kinds

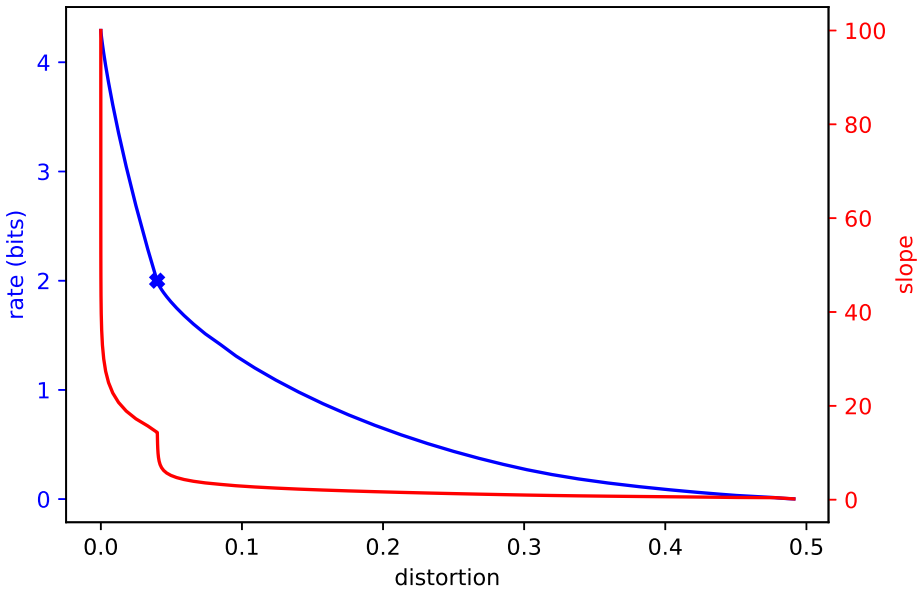


Figure 5: A sweet spot in the rate-distortion function

the world of relevance to the information-processing system, even through a very severe, 2 bit bottleneck. I claim that this is no coincidence. Our representation-attributing practices gravitate towards this kind of situations.

To see how sweet spots in rate-distortion curves and representations are related, consider now what an optimal encoding-decoding strategy would look like. That is, how should the encoder encode the information coming from the source, and how should the decoder decode the signal coming from the encoder, so that the resulting expected distortion between original and decoded signal is the minimum achievable, at the sweet spot?

Optimal Encoding Strategy: First divide the incoming signal in two halves, one corresponding to properties P_1 through P_5; the other corresponding to properties P_6 through P_{10}.

If there is a majority of 1s in the first half of the original signal set the first bit of the signal to 1. Otherwise set it to 0. Ditto for the second half of the original signal and the second bit of the signal.

Optimal Decoding Strategy: If the first bit in the incoming signal is 1, set the first half of the decoded signal to 11111. Otherwise, set it to 00000. Ditto for the second bit and the second half of the decoded signal.

How should we interpret what encoder and decoder are doing here? A natural way is this: they are using the presence or absence of properties in an HPC cluster as diagnostic of the presence or absence of the underlying natural kind—this would be the encoding part—and then taking the resulting signals as representing the presence of a paradigmatic instance of the kind, one that has all the properties in the cluster—this would be the decoding part. HPC kinds being what they are, frequently the first half of the incoming signal will resemble the paradigmatic presence of the first kind (11111) or its paradigmatic absence (00000), and the same will happen with the second half and the second kind. That is why this encoding-decoding strategy works so well.

In describing this optimal strategy I have helped myself to representational vocabulary; it has been useful in order to explain how the strategy works, and how come that behaving in this particular way achieves low distortion at low rates: it is because each of the two bits in the signal is caused by, and causes, behavior that is optimally attuned to the probabilistic structure of each of the two natural kinds in the model world, respectively. Nothing going on in this system falls outside the purview of Shannonian information theory—of information theory *tout court*, so at least in this kind of cases representational talk depends on no non-informational fact.

We can now understand better what's lacking in the philosopher of mind's information-theoretic tookit: it is entirely possible, and computationally trivial, to calculate, e.g., Skyrms's pmi between each of the possible signals (00, 01, 10 and 11) and each of the possible world states (all 1024 of them, from 0000000000 to 1111111111). Doing so would leave us with 4 vectors (one for each signal) with 1024 entries each (one for each world state.) First, this is an unwieldy collection of numbers, which doesn't bring out the relevant structure. For example, if the probability of children nodes being *on* conditional on their parent being *on* was .96 instead of .95 the rate-distortion curve

Seattle, WA; 1-4 November 2018 -455-

would be qualitatively identical, with a sweet spot in exactly the same place, yet most numbers in the Skyrmsian informational content vectors would change. Second, and most important, nothing in those 4096 numbers allows us to infer the presence of a sweet spot. The relevant information is simply not there, depending as it does on a distortion measure which is not used in computing Skyrmsian informational contents.

If this is approximately right, the question about what makes representational talk explanatory is readily answered: saying that a certain vehicle is a representation conveys something quite specific about its informational context. It says that the vehicle is part of an encoding-decoding strategy that exploits a sweet spot in a rate-distortion curve—where the curve is in turn fixed by the probabilistic structure of the world, and the target distortion measure. This, in less technical terms, translates to saying that the vehicle is summarizing *relevant* (this is where the distortion measure comes in) aspects of the current situation in an optimal, if lossy, manner, made possible by *how the world* is (this is where the probabilistic structure of the world comes in.) This explication of the explanatory contribution of representations can be turned into an explicit answer to what makes something a representation—an answer, that is, to what Artiga (2016) calls the metasemantic question.

The Rate-Distortion Approach: A signal, S, in a certain information-processing pipeline, P, is a representation if the following two conditions are met:

Existence: There are sweet spots in the rate-distortion curve associated with P.

Optimality: S is produced as part of an encoder-decoder strategy that occupies the vicinity of one of these sweet spots.

So, *pace* Dretske, the core information-theoretic notions of entropy, rate, distortion, etc. *can* provide invaluable insight into the representational status of individual signals. If the rate-distortion approach is on the right track, those information-theoretic notions, through the existence condition, specify the kind of setup where representations live, which then the optimality condition can use to provide a criterion for the representational status of individual signals.

I offer the foregoing discussion as a preliminary case for the rate-distortion approach to representation: it shows how postulating representations is explanatory, even if these representations depend just on (Shannon) information. It illuminates the difference in representational status between cue-driven examples, such as Schulte's vasopressin; and vervet alarm calls, and other similar examples. To complete my case I now show how the ways to bridge the gap between natural and nonnatural information discussed in section 2 can be seen as unwitting attempts to get at rate-distortion sweet spots.

4 There is no Gap to Bridge

What does it take for the existence condition to be met? That is to say, what circumstances result in sudden drops in the slope of the rate-distortion curve? We have seen one such family of circumstances: if the pattern in which properties are instantiated

in the source is noisily replicated in a cluster then sudden drops are to be expected: distortion will decrease with rate up to the point where all the main sources of variation in property instantiations are accounted for, and all that remains is the residual noise in instantiations within each cluster. Take a look again at figs. 4 and 5: to describe this source we basically need enough rate to account for the two main sources of variation: P_1 and P_6. This is not all there is to the world, because it's possible for the other properties to (fail to) token independently of their parent, but the unlikeliness of these departures makes the extra rate comparatively less useful.

Noisy replication of property instantiations is at the core of the HPC theory of natural kinds, as we saw above. This means that, in general, the presence of HPC natural kinds in a source will create sweet spots. This opens a line of argument in favor of reference magnetism from information-theoretic premises: reference magneticsm should be seen as making a point about the kind of probabilistic structure that an information-processing pipeline must be attuned to, if signals are to effect the kind of optimal lossy compression that underlies our representation-attributing practices. Reference magnetism is just a way of meeting part of the existence condition.

Regarding the suggestion, by Sterelny, Burge and others, that representations inhere preferably on signals sitting in a one-to-many-to-one pipeline, I submit that the many-to-one aspect of this suggestion aims at meeting the optimality condition; the one-to-many aspect, together with reference magnetism, aims at meeting the existence condition.

The first thing to note here is that the *Optimal Encoding Strategy* presented above enforces what Sterelny calls robust tracking and Burge calls constancy: the strategy consists in considering all properties coming from each of the two clusters and setting the relevant bit to 1 only if a majority of those properties are instantiated. That is, the encoder is taking a multiplicity of configurations (e.g., the first half of the incoming signal being 00111, 01011, 10111, etc.) to a single output: the first bit of the signal being 1. Furthermore, that part of the signal will be decoded as 11111: from there on, the system downstream will treat whatever is out there in the world as a paradigmatic member of the first kind. The system is recovering the presence of a natural kind out of many different, noisy instantiation patterns. This is a clear instance of constancy. Suppose that the encoder, insted of being many-to-one, depended on a single cue; say, suppose it set the first bit to 1 if one of the children properties (say, P_2) was instantiated, and to 0 otherwise. In such a cue-driven setup, the best encoder-decoder arrangement possible is marked by the blue circle in fig. 6. This has double the distortion than than the optimal encoding (marked by the blue cross) which sits right on top of the optimal rate-distortion curve. This cue-driven system would not meet the optimality condition, which means that a many-to-one architecture is instrumental to meeting it.

Finally, the target distortion measure in the information-processing pipeline can be seen as that which Sterelny's one-to-many condition on representation is actually tracking. Using, for example, the Hamming distance as a distortion measure is tantamount to assuming that all of the properties of the natural kinds are relevant for downstream processing. One natural way in which this may happen is when the agent is to respond flexibly to the presence of the natural kind: in different contexts or states different properties of the kind might be relevant and, for example, the presence of a tree might

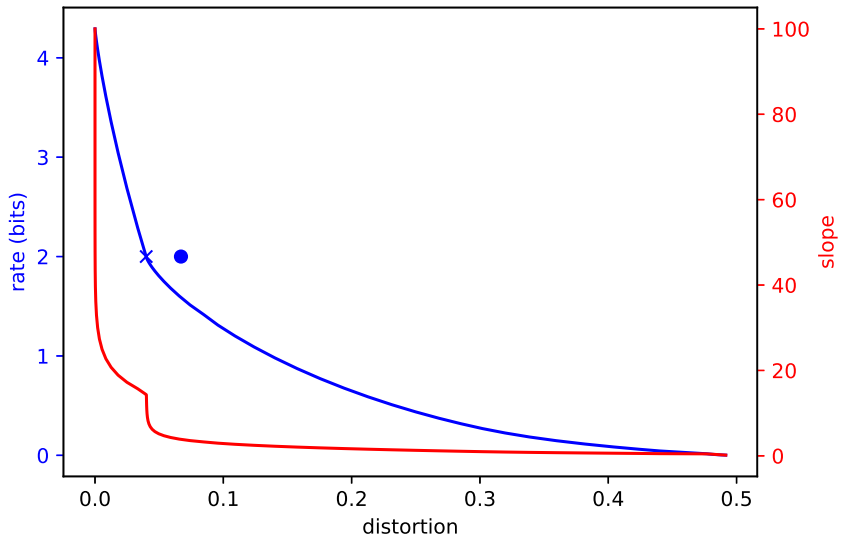


Figure 6: Cue-driven encoding

be sometimes relevant to behavior because it bears fruit (if the agent is hungry) and some other times because it has a dense cover (if the agent is looking for shelter.)

Caring about all (or many) properties of the kind is what makes the rate-distortion curve display a sweet spot. If, instead, the agent has a rigid, stereotyped response to the presence of members of the kinds—that is, if it only cares about the presence of one property, which is the property that makes that rigid behavioral response fitness-conducive, then the curve is as presented in fig. 7. Rigid behavioral responses make the probabilistic structure of the kinds largely irrelevant. As a result, the system behaves as if a coin were tossed, where heads would mean that the target property is tokened, and tails that it is not. This arrangement does not meet the existence condition. Sterelny's broad-banded responses are, again, a way of getting at rate-distortion sweet spots.

References

Arimoto, S 1972, 'An algorithm for computing the capacity of arbitrary discrete memoryless channels', *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20.

Artiga, M forthcoming, 'Beyond Black Spots and Nutritious Things: A Solution to the Indeterminacy Problem', *Dialectica*.

Artiga, M 2016, 'Liberal Representationalism: A Deflationist Defense', *dialectica*, vol.

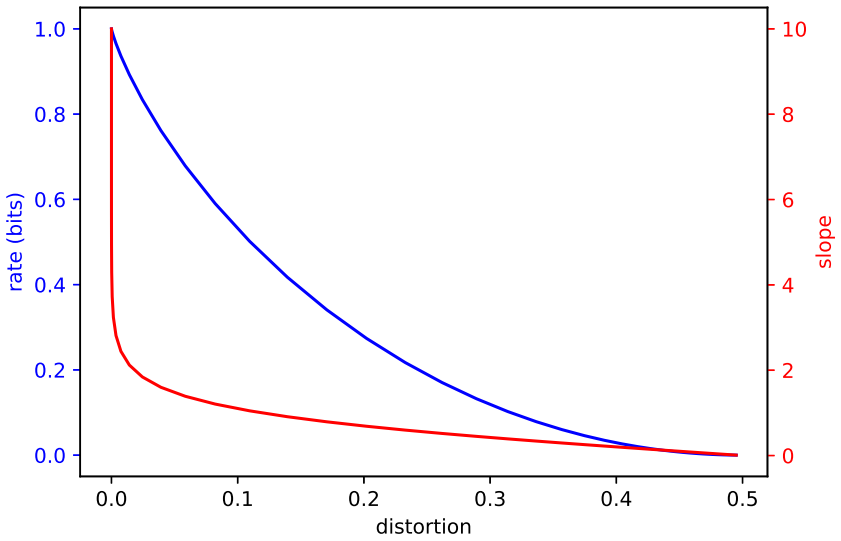


Figure 7: Rigid behavioral response

70, no. 3, pp. 407–430.

Blahut, R 1972, 'Computation of channel capacity and rate-distortion functions', *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473.

Boyd, R 1989, 'What Realism Implies and What It Does Not', *Dialectica*, vol. 43, no. 1-2, pp. 5–29.

Burge, T 2010, *Origins of objectivity*, Oxford University Press.

Cover, TM & Thomas, JA 2006, *Elements of Information Theory*, New York: Wiley.

Dretske, F 1981, *Knowledge and the Flow of Information*, The MIT Press.

Lewis, D 1983, 'New work for a theory of universals', *Australasian journal of Philosophy*, vol. 61, no. 4, pp. 343–377.

Lewis, D 1984, 'Putnam's paradox', *Australasian Journal of Philosophy*, vol. 62, no. 3, pp. 221–236.

Mann, SF 2018, 'Consequences of a Functional Account of Information', *Review of Philosophy and Psychology*, pp. 1–19.

Martínez, M 2013, 'Teleosemantics and Indeterminacy', *Dialectica*, vol. 67, no. 4, pp.

427–453.

Millikan, R 1984, *Language, Thought and Other Biological Categories*, The MIT Press.

Papineau, D 1987, *Reality and Representation*, Basil Blackwell.

Rathkopf, C 2017, 'Neural information and the problem of objectivity', *Biology & Philosophy*, vol. 32, no. 3, pp. 321–336.

Ryder, D 2006, 'On Thinking of Kinds', in G Macdonald & D Papineau (eds), *Teleosemantics*, Oxford University Press, pp. 1–22.

Ryder, D 2004, 'SINBAD Neurosemantics: A Theory of Mental Representation', *Mind & Language*, vol. 19, no. 2, pp. 211–240.

Scarantino, A 2015, 'Information as a probabilistic difference maker', *Australasian Journal of Philosophy*, vol. 93, no. 3, pp. 419–443.

Schulte, P 2015, 'Perceptual representations: A teleosemantic answer to the breadth-of-application problem', *Biology & Philosophy*, vol. 30, no. 1, pp. 119–136.

Seyfarth, RM, Cheney, DL & Marler, P 1980a, 'Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication', *Science*, vol. 210, no. 4471, pp. 801–803.

Seyfarth, RM, Cheney, DL & Marler, P 1980b, 'Vervet monkey alarm calls: Semantic communication in a free-ranging primate', *Animal Behaviour*, vol. 28, no. 4, pp. 1070–1094.

Sider, T 2014, *Writing the Book of the World*, Reprint edition., Oxford University Press, Oxford.

Skyrms, B 2010, *Signals: Evolution, Learning & Information*, New York: Oxford University Press.

Sterelny, K 2003, *Thought In A Hostile World: The Evolution of Human Cognition*, John Wiley & Sons, Malden, MA.

The Proportionality of Common Sense Causal Claims

Jennifer McDonald

This paper defends strong proportionality against what I take to be its principal objection – that proportionality fails to preserve common sense causal intuitions – by articulating independently plausible constraints on representing causal situations. I first assume the interventionist formulation of proportionality, following Woodward.[1] This views proportionality as a relational constraint on variable selection in causal modeling that requires that changes in the cause variable line up with those in the effect variable. I then argue that the principal objection derives from a failure to recognize two constraints on variable selection presupposed by interventionism: *exhaustivity* and *exclusivity*.

[1] Woodward 2003

1. Introduction

Yablo's principle of proportionality holds, roughly, that something counts as a cause of some effect just in case it includes the appropriate degree of causal information.[2] Proportionality has been put to various philosophical uses, such as a proposed solution for the causal exclusion argument, and as a justification and explanation of the dependence on high-level causal explanations in the special sciences. However, the precise formulation of such a principle has proven to be controversial.

I take the most promising formulation to be an interventionist one, following Woodward.[3] Such a formulation defines proportionality as a relational constraint on variable selection in causal modeling. In this paper, I argue that this formulation works well as it is – contra Franklin-Hall (see 2016) – so long as we recognize two independently plausible background requirements on variable selection. I call these *exhaustivity* and *exclusivity*. Exhaustivity holds that a variable must take at least one of its values. Exclusivity holds that a variable can take at most one of its values. Both constraints are relative to, and thereby help to make explicit, the modal assumptions implicit in causal inquiry.

Finally, with these requirements in place, I defend proportionality against its principal objection: that it fails to preserve fundamental causal intuitions. I demonstrate how this concern derives from a failure to recognize and integrate the modal assumptions implicit in causal inquiry, in tandem with an inappropriate use of variables to represent causal situations.

2. Interventionism

The formulation of proportionality that I endorse comes directly from Woodward, and is defined in terms of his interventionist account of causation. Interventionism expands on the intuition that causal claims provide

[2] Yablo 1992
[3] Woodward 2003, 2008a, 2008b, 2010, 2016

manipulability information. If *X* causes *Y*, then manipulating or changing *X* is a way of manipulating *Y*. It then exploits the language of causal models to identify and articulate different causal relations of interest. A causal model can take a variety of forms, such as graphical, potential-outcome, and structural-equations models.[4] However, I'll restrict discussion of causal models in this paper to graphical models. A graphical model is, essentially, a set of variables – representing the causal relata – and a directed binary relation between them – representing causal influence.

Interventionism then defines the notion of an *intervention* on a system. An intervention, *I*, first must directly change the value of some variable, *X*, in such a way that it breaks the dependence that *X* may have had on other variables in the system. Second, *I* must be designed in such a way that any change in the effect variable, *Y*, will be the direct result of *X* and not of *I* itself. Finally, *I* must be wholly independent of other possible causes of *Y*, whether such causes are represented by the given model or not. A more precise formulation than this won't matter for the purposes of this paper.[5]

With this in place, the interventionist then defines a basic notion of cause, which corresponds most closely with the intuitive notion of *causal relevance*:

> (Principle M) *X* causes *Y* iff there are background circumstances B such that if some (single) intervention that changes the value of *X* (and no other variable) were to occur in B, then Y would change. (Woodward 2003, 222)

That is, in order for *X* to be a cause of *Y*, the change in *X* from one value to another as the result of an intervention corresponds to the change in *Y* from one value to another, given some fixed set of background parameters. Various kinds of causal relations are then captured by refinements on this basic notion. Due to

[4] See Greenland and Brumback 2002 and Hitchcock 2009 for overviews of causal models.
[5] See Woodward 2003, chapter 3, especially 98

the irrelevance of these and further details to my argument, I'll leave my overview of interventionism here.[6]

3. Proportionality as Relational Constraint on Variable Selection

Interventionism places variables front and center in how we represent and inquire into causation. Thus, more needs to be said about the criteria for variable selection. Although the variables can be taken to represent different things, I will assume throughout that the set of values of a particular variable represents a set of properties – constrained by a given property type – that are possibly instantiated by some particular thing. The assumed causal relata of this paper will therefore be property instantiations.

This paper addresses two questions relevant to variable selection: (i) What determines the range of values that a variable can take? (ii) At what level of description should the values of the variables be? Proportionality has been proposed as an answer to (ii). However, after laying out the proposal, I'll go on to argue that while (ii) can be answered by the principle of proportionality, it can only do so alongside an appropriate answer to (i). One aspect of such an answer is that the background modal context determines the range of values that a variable takes.

Constraints on variable selection can be divided into two kinds: relational constraints and non-relational constraints. *Relational constraints* pertain to the extrinsic nature of the variables in a causal model, to how "variables relate to one another." (Woodward 2016, 1056) One example of such a constraint is stability.[7] *Stability* is the persistence of the causal relation between a cause variable and an effect variable, despite changes in the background conditions. The more changes such a relation can survive, the more stable it is.

[6] See Woodward 2003, chapter 2, especially section 3
[7] See Woodward 2010, 2016

Proportionality is just such a relational constraint. It holds that changes in a cause variable should line up with changes in an effect variable. Intuitively,

> Proportionality has to do with whether changes in the state of the cause 'line up' in the right way with changes in the state of the effect and with whether the cause and effect are characterized in a way that contains irrelevant detail. (Woodward 2010, 287)

Take Yablo's pigeon example.[8] Sophie the pigeon is trained to peck at red things and only at red things. She then pecks at a paint chip, which is a particular shade of red – scarlet. Which of the following is causally relevant to Sophie's pecking: the chip's being red or the chip's being scarlet?

When translated into interventionist terms, this becomes a false dichotomy. Take the variable, *P*, to be a variable representing whether the pigeon pecks or not. It can take the values: {*peck*, *not-peck*}. Now consider two alternative variables for representing the property-instantiations of the paint chip: the variable, *R*, which can take the values {*red*, *not-red*}, and the variable, *T*, which can take the values {*taupe, scarlet, cyan, mauve, crimson,* etc.}, where 'etc.' stands for all other physically possible colors at the same grain as those already made explicit. According to Principle M, the causal model in which *R* stands as causally relevant to *P* is just as accurate as one in which *T* so stands. In the *R* model, *R* is causally relevant to *P* because an intervention on *R* that changes its value from *not-red* to *red* changes *P*'s value from *not-peck* to *peck.* In the *T* model, *T* is causally relevant to *P* because an intervention on *T* that changes its value from *taupe* to *scarlet* changes *P*'s value from *not-peck* to *peck*.

Interventionism therefore doesn't ask the question, which variable stands in a causal relation to *P*? For, the answer is 'both'. *R* and *T* are each causally relevant to *P*. But, this doesn't mean that their respective relationship to *P* is the same. *R* is *proportional* to *P*, while *T* is not. All of the changes in *R* line up with changes in *P* – every intervention on *R* corresponds to a change in *P*. But only some of the

[8] Yablo 1992

changes in *T* line up with those in *P* – only certain interventions on *T* correspond to changes in *P*. The intervention that changes the value of *T* from *taupe* to *cyan*, for example, will not change the value of *P*.

Woodward defines proportionality more explicitly as,

> (P) There is a pattern of systematic counterfactual dependence (with the dependence understood along interventionist lines) between different possible states of the cause and the different possible states of the effect, where this pattern of dependence at least approximates to the following ideal: [it] should be such that (a) it explicitly or implicitly conveys accurate information about the conditions under which alternative states of the effect will be realized and (b) it conveys only such information – that is, the cause is not characterized in such a way that alternative states of it fail to be associated with changes in the effect. (2010, 298)

There are two views on what this difference between variables like *R* and *T* means. The first takes proportional variables such as *R* to represent genuine causes, while non-proportional variables such as *T* represent merely causally relevant factors. Proportionality is thereby considered a necessary constraint on causation. Call this *strong proportionality*.[9] The second view takes proportionality to be a merely pragmatic constraint on causal explanation.[10] Call this *weak proportionality*. Throughout this paper, I assume and defend strong proportionality.

4. Non-Relational Constraints: Exhaustivity and Exclusivity

Non-relational constraints, on the other hand, pertain to the intrinsic nature of the variables in a causal model. These constraints "can be applied to variables, individually, independently of how they relate to other variables." (Woodward

[9] See List and Menzies 2009; Menzies and List 2010; and Papineau 2013
[10] See Woodward 2015; Shapiro and Sober 2012; McDonnell 2017; and Weslake 2013, 2017

2016, 1057) One example is *metaphysical naturalness*, which requires that variables pick out only natural properties, on some understanding of 'natural'.[11]

What I propose to call the exhaustivity and the exclusivity constraint are similarly non-relational constraints. Take exhaustivity first. The *exhaustivity constraint* requires that a variable's values capture the entire range of relevant possibilities for whatever type of thing the variable represents. An exhaustive variable is one that must take one of its values, given whatever background modal constraints are in place.

Since I've restricted this discussion to variables whose values represent the property instantiation of some target object, I can define exhaustivity in more precise terms. *Exhaustivity* is the constraint on a variable in a causal model that holds that its values must jointly represent the range of possibilities of property instantiation by the given object for the given property-type. If the property-type is a color, for example, then the values must somehow exhaust the color spectrum. This can be done quite simply with a binary variable that can take the values: {*some particular color*, *not-(that particular color)*}.

Next, the *exclusivity constraint* holds that the values of a given variable should be such that any one excludes all the others. Woodward references exclusivity when he writes,

> When considering the values of a single variable, we want those values to be logically exclusive, in the sense that variable *X*'s taking value *v* excludes *X*'s also taking value *v'*, where $v \neq v'$. (2016, 1064)

In other words, if two things are not exclusive – if they could occur together – then they should be represented by distinct variables. While exhaustivity holds that a variable should take *at least* one of its values, exclusivity holds that a variable should take *at most* one of its values.

[11] See Lewis 1983; Menzies 1996; Paul 2000; and Franklin-Hall 2016

Importantly, exhaustivity and exclusivity are each relative to a background modal context. In possible worlds terminology, the modal context is the set of possible worlds relevant to the truth of the counterfactual that captures the causal claim. It can be described as a set of worlds, or perhaps more succinctly as a list of background assumptions that define such a set. These assumptions can include any constraint that operates in a law-like fashion.

For example, the causal claim, "The chip's being scarlet caused the pigeon to peck," corresponds to the counterfactual, "Had the chip not been scarlet, the pigeon wouldn't have pecked." The modal context of this claim and corresponding counterfactual is the set of possible worlds that determines whether the counterfactual is true. So, if this claim and counterfactual are meant to represent a *specific* causal situation near a local paint chip factory that specializes in just the colors scarlet and cyan, and no others, then the relevant set of possible worlds will be constrained to those in which the paint chip takes one of the two factory colors – cyan or scarlet. In this context, the variable, *C,* that can take the values {*cyan, scarlet*}, is an exhaustive variable. Further, given this set of worlds, the counterfactual is true.

If instead these are meant to represent any *general* causal situation involving paint chips and a red-pecking pigeon, then the relevant set of possible worlds will be more inclusive, including all worlds in which the paint chip takes any color within the color spectrum. *C* is not exhaustive relative to this more inclusive modal context. But the variable *T*, from before, is. Given this more inclusive set of worlds, the counterfactual is false, since the pigeon will peck in response to shades of red other than scarlet.

A point of note here is that the constraints of exhaustivity and exclusivity are indeed non-relational constraints in the sense previously defined. Although they are relative to the modal context, they are *not* relative to other variables in the model. They are properties of a variable taken independently as a representation of the target scenario.

I hold that causal models successfully represent causal situations in part by requiring exhaustive and exclusive variables. Proportionality, defined in terms of causal models, also requires exhaustive and exclusive variables. A significant upshot of this is that the proportional cause is not only relative to the target effect variable, but also to the background modal context.

5. Interventionist Proportionality Does the Trick

Franklin-Hall contends that Woodward's formulation of proportionality doesn't successfully prioritize intuitively proportional causal relata, such as red in the pigeon example. However, as I'll argue, presupposing my notion of exhaustivity corrects for this objection.

Franklin-Hall argues that proportionality as laid out in section 3 is inadequate for capturing the kind of causal explanation we're looking for. To do so, she calls upon Sophie and her paint chip. She then introduces a comparison between the causal variable, *R*, that can take the values: {*red*, *not-red*}, (as above), and a variable, *C*, that can instead take the values: {*cyan, scarlet*} (as above). *R*, as before, is proportional to, and therefore a genuine cause of, *Y*. But, she argues, *C*, too, is proportional to *Y*, since every possible intervention on *C* changes the value of *Y*. An intervention on *C* that changes its value from *cyan* to *scarlet* changes *Y* from *not-peck* to *peck*, and an intervention that changes *C*'s value from *scarlet* to *cyan* changes *Y*'s value from *peck* to *not-peck*. Thus, the changes in *C* line up with the changes in *Y* just as well as the changes in *R* do. The problem, then, is that proportionality, as formulated, is insufficient to its intended task. It fails to privilege a variable like *R* over one like *C*, and so fails to prioritize a causal model that uses *R* over one that uses *C*.

In response to this problem, a natural move would be to find a way to disqualify variables like *C* from the arena. Intuitively, *C* is not the right kind of variable. But, why not? I propose that our aversion to variables like *C* is due to their failure to exhaustively represent the implicit modal context of the situation. The background possibilities relative to the paint chip include the full color spectrum.

Unless the possible color of the paint chip is restricted in some way – by the local factory, for example – then the target object can fail to take one of *C*'s two values. There are other physically possible colors that the paint chip could have – such as beige or olive green – and *C*'s values fail to represent these possibilities.

Relative to the implicit modal context, then, *C* is not an exhaustive variable. The variable, *R*, on the other hand, is exhaustive, since the object must take one of *R*'s two values. By requiring exhaustive variables, *C* is discounted as a candidate variable *relative to the implicit modal context*, and *R* takes privilege as the proportional cause.

In general, two variables are in proper competition with each other over which is proportional to some effect variable only when they are exhaustive relative to the same modal context. *C* and *R* are not competitors for proportionality relative to *Y*, since only one of them can contain an exhaustive set of active possibilities relative to any given modal context.

6. Preserving Causal Intuitions

The strongest objection to proportionality, as raised by Bontly, Shapiro and Sober, McDonnell, and Weslake, is that it seems to render many common sense causal claims false.[12] Call this the *objection from common sense*. It objects to strong proportionality by attempting to demonstrate that if proportionality is required of something to be a cause, then many things that we would naturally call causes don't actually qualify.

Take as an example the situation where Socrates drinks hemlock and then dies, and the corresponding causal claim, 'Socrates's drinking hemlock caused him to die'. The objection goes that drinking hemlock is not actually proportional to Socrates dying. For example, if Socrates had not drank hemlock, but still consumed it – by eating a dozen leaves, for example – then he still would have

[12] See Bontly 2005; Shapiro and Sober 2012; McDonnell 2017; and Weslake 2013, 2017

died. This seems to show that the changes in the variable that represents Socrates drinking hemlock don't line up with the changes in the variable that represents Socrates dying. The first variable could change values from *Socrates-drinks-hemlock* to *Socrates-eats-hemlock* and the second variable would retain the value *Socrates-dies*. This common sense causal claim is therefore not proportional. The proportional cause should be, instead, *consuming hemlock*.

However, this objection is mistaken. It fails to respect the exhaustivity constraint on variable selection, and thereby equivocates between different background modal contexts. It further fails to respect exclusivity, and thereby runs together what should be different variables. Rectifying this illuminates the implicit proportionality of common sense causal claims.

First, the objection ignores the fact that proportionality, in requiring exhaustive and exclusive variables, is relative to modal context. Take the hemlock example just outlined. Importantly, this example and corresponding claim are under-defined.[13] Translated into interventionist terms, all that this description provides is that there is some variable that takes a value that represents Socrates drinking hemlock, and an intervention on this variable changes the value of some other variable to one that represents Socrates dying. But, a number of different variables could represent the purported cause, and a number of different models could represent its relationship to the effect of Socrates' dying. Which of these is accurate depends on what the relevant alternatives to drinking hemlock are. How these details get filled in will determine whether or not the variable that represents Socrates drinking hemlock is proportional.

I hold that the common sense claim that drinking hemlock causes Socrates's death implicitly takes the relevant alternative to be Socrates's *not* drinking hemlock. The default context is taken to be that hemlock was the only possible poison, and drinking it the only possible means of consumption. Given this context, the exhaustive variable would take the values {*drinks-hemlock*, *doesn't-*

[13] I take this to be common knowledge. See Franklin-Hall 2016; McDonnell 2017; and Weslake 2017

drink-hemlock}. But, such a variable is indeed proportional to the effect variable. Thus, the common sense cause is, in fact, proportional.

Such a defense requires that common sense claims be implicitly relative to a modal context. I'm not the first to relativize common sense claims to context. Philosophers such as Mackie and Schaffer make such a move, albeit with different ends in mind.[14] However, both McDonnell and Weslake explicitly deny this kind of relativity.[15] They claim that the very fact that we have strong and convergent intuitions about common sense examples, despite their being under-determined, demonstrates that the intuitions are not sensitive to filling in details.

In response, I argue that we respond to common sense causal examples in the same way that we respond to standard conversations. According to Grice, communication is governed by a set of conversational maxims.[16, 17] The maxims most relevant to how an audience engages with these under-defined causal examples are the maxims of *quantity* and *relation*. Taken together, these maxims enjoin an interlocutor to,

> Make your contribution as informative as is required (for the current purposes of exchange)….[and no] more informative than is required,….[and b]e relevant. (1989, 26 – 27)

Thus, the conversationally natural way to fill in the modal context of these examples is to take each fact as informative and relevant, and to assume that all informative facts have been provided.

The only information provided by the hemlock example is the following: (i) Socrates drinks hemlock. (ii) Socrates dies. The Gricean maxims tell us that this is all the information needed, and that nothing significant has been left out. So, the details are filled in as continuous with everyday life. In possible world speak,

[14] See Mackie 1974, especially chapter 2; and Schaffer 2005
[15] McDonnell 2017; Weslake 2017
[16] See Grice 1989
[17] Bontly makes a similar point (see 2005)

we're looking only at worlds which have a similar environment, a biologically similar Socrates, etc., and in which laws of metaphysical necessity hold.

The causal focus is on Socrates's drinking hemlock. This means that in evaluating the causal relationship, everything else is held fixed and the fact of the drinking hemlock is varied. Due to the absence of any other details, the only real alternative to Socrates's drinking hemlock is his not drinking hemlock. Nothing suggests that there are alternative means of consuming the hemlock. Further, it's not a common occurrence in everyday life to have alternative means of consuming a given poison. Treating *eating hemlock* as a relevant alternative would be to arbitrarily introduce something that wasn't otherwise specified, and whose presence can't be justified by everyday experience.

The objection from common sense assumes different possible alternatives than what I take to be implicit, and then tries to say that relative to these other alternatives, the common sense causal claim is not proportional. I have argued that the common sense cause is simply not relative to these other alternatives.

However, even given other possible alternatives, the common sense cause would still be proportional. The second mistake that the objection makes is that it fails to appreciate the constraint of exclusivity.

The objection holds that there is some relevant alternative to Socrates's drinking hemlock that preserves his consuming it. Take as an arbitrary alternative his eating hemlock. Socrates could both drink and eat the hemlock – he could wash down a hemlock salad with a glass of hemlock milk, for example. Following exclusivity, then, these possibilities should be represented by distinct variables – one that can take the value *drinks-hemlock*, call this *D*, and one that can take *eats-hemlock*, call this *E*.

But, now there is no problem. Following Woodward's response to early pre-emption cases,[18] we can hold *E* fixed at the value that represents Socrates not eating the hemlock, and see if the changes in *D* – which we can ensure meets exhaustivity by giving it the second value *doesn't-drink-hemlock* – line up with the changes in the effect variable. They do. When an intervention sets the value of the cause variable to *drinks-hemlock*, the effect variable takes the value *dies*. When an intervention sets the value of the cause variable instead to *doesn't-drink-hemlock*, the effect variable changes value to *doesn't-die*. Once again, the common sense cause is proportional.

If, on the other hand, the situation is such that Socrates's drinking hemlock is indeed mutually exclusive with his eating hemlock, then *drinks-hemlock* and *eats-hemlock* could be values of the same variable. Imagine that Socrates's jailor only has enough money to purchase either hemlock leaves or hemlock milk, but not both. In this case, neither Socrates's drinking nor his eating will be proportional. The proportional cause is instead his consuming hemlock. The proportional variable will therefore be one that takes as values {*consumes-hemlock*, *doesn't-consume-hemlock*}.

But, this is not in conflict with common sense – so long as we abstract away from normal everyday circumstances, and instead genuinely fix the situation as one in which Socrates is forced to consume hemlock, arbitrarily receiving hemlock leaves or milk. When, given this background, we're asked what causes Socrates's death, it is natural to say that it was his consuming hemlock. After all, it isn't the drinking nor the eating that makes a difference to whether Socrates dies, since had he not done one he would have done the other. It is his consuming hemlock rather than not.

Finally, I'd like to point out that the intuition that Socrates's consuming hemlock is the more proportional cause is actually misguided. The naïve intuition holds that an exhaustive and exclusive variable with the value *consumes-hemlock* – call this H_1 – is more proportional to the exhaustive and exclusive variable with the

[18] See Woodward 2003

value *drinks-hemlock* – call this H_2. But, the modal context to which H_1 will be exhaustive is different than that to which H_2 will be. They're therefore not even in competition for proportionality. Instead, I suggest that this intuition is a response to the fact that H_1's modal context is more inclusive than that of H_2. H_1 can accurately (and proportionally) represent the cause of Socrates's death in a wider range of situations than can H_2. But, this is about stability – as earlier defined – not about proportionality. The model that employs H_1 is simply *more stable* than that which employs H_2. This putative proportionality intuition is actually responding to the property of stability.

7. Conclusion

In this paper, I have defended the interventionist formulation of proportionality by explicating the exhaustivity and exclusivity constraints, and stipulating that proportionality requires variables that meet these constraints.

These constraints have been defined on the assumption that a variable represents a particular object's instantiations of a particular type of property. But, they are easily generalized to cover alternate objects of representation. Take events, for example. If variables represent particular kinds of events occurring or failing to occur, then exhaustivity would require that the values of a variable cover the entire range of possibilities of event occurrence for whatever type of event the variable represents. Exclusivity would require that the values of a variable be event occurrences such that no two could occur simultaneously.

Finally, I have articulated how the interventionist formulation of proportionality responds to the objection from common sense. Such an objection dissolves once the explicated constraints on variable selection are honored.

8. References

Bontly, Thomas. 2005. "Proportionality, Causation, and Exclusion." *Philosophia* 32 (1-4): 331 – 48

Franklin-Hall, Laura. 2016. "High-Level Explanation and the Interventionist's 'Variables Problem.'" *British Journal for the Philosophy of Science* 67 (2):553 – 77

Greenland, Sander, and Babette Brumback. 2002. "An Overview of Relations Among Causal Modelling Methods." *International Journal of Epidemiology* 31:1030 – 37

Grice, H. Paul. 1989. *Studies in the Way of Words.* Cambridge, MA: Harvard University Press

Hitchcock, Christopher. 1996. "The Role of Contrast in Causal and Explanatory Claims." *Synthese* 107 (3): 395 – 419

--- 2009. "Causal Modelling." In *The Oxford Handbook of Causation*, ed. Helen Beebee, Christopher Hitchcock, and Peter Menzies, 299 – 314. Oxford: Oxford University Press

Lewis, David. 1983 "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343 – 77

List, Christian, and Peter Menzies. 2009. "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106 (9): 475 – 502

Mackie, J. L. 1974. *The Cement of the Universe.* Oxford: Oxford University Press

McDonnell, Neil. 2017. "Causal Exclusion and the Limits of Proportionality." *Philosophical Studies* 174 (6): 1459 – 74

Menzies, Peter. 1996. "Probabilistic Causation and the Pre-emption Problem." *Mind* 105 (417): 85 – 117

Menzies, Peter, and Christian List. 2010. "The Causal Autonomy of the Special Sciences." In *Emergence in Mind*, ed. Cynthia Macdonald and Graham Macdonald, 108 – 28. Oxford: Oxford University Press

Papineau, David. 2013. "Causation is Macroscopic but not Irreducible." In *Mental Causation and Ontology*, ed. Sophie C. Gibb and Rögnvaldur Ingthorsson, 126 – 52. Oxford: Oxford University Press

Paul, L.A. 2000. "Aspect Causation." *The Journal of Philosophy* 97 (4): 235 – 56

Schaffer, Jonathan. 2005. "Contrastive Causation." *The Philosophical Review* 114 (3): 297 – 328

Shapiro, Larry, and Elliott Sober. 2012. "Against Proportionality." *Analysis* 72 (1): 89 – 93

Weslake, Bradley. 2013. "Proportionality, Contrast, and Explanation." *Australasian Journal of Philosophy* 91 (4): 785 – 97

--- 2017. "Difference-Making, Closure, and Exclusion." In *Making a Difference,* ed. Helen Beebee, Christopher Hitchcock, and Huw Price, 215 – 32. New York: Oxford University Press

Woodward, James. 2003. *Making Things Happen.* Oxford: Oxford University Press

--- 2008a. "Mental Causation and Neural Mechanisms." In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, ed. Jakob Hohwy & Jesper Kallestrup, 218 – 62. Oxford: Oxford University Press

--- 2008b. “Response to Strevens.” *Philosophy and Phenomenological Research* 78 (1): 193 – 212

--- 2010. “Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation.” *Biological Philosophy* 25 (3): 287 – 318

--- 2015. “Interventionism and Causal Exclusion.” *Philosophy and Phenomenological Research* 91 (2): 303 – 47

--- 2016. “The Problem of Variable Choice” *Synthese* 193 (4): 1047 – 72

Yablo, Stephen. 1992. “Mental Causation” The Philosophical Review 101 (2): 245 – 80

Species as Models

Abstract: This paper argues that biological species should be construed as abstract models, rather than biological or even tangible entities. Various (phenetic, cladistic, biological etc.) species concepts are defined as set-theoretic models of formal theories, and their logical connections are illustrated. In this view organisms relate to a species not as instantiations, members, or mereological parts, but rather as phenomena to be represented by the model/species. This sheds new light on the long-standing problems of species and suggests their connection to broader philosophical topics such as model selection, scientific representation, and scientific realism.

1 Introduction

Biological species has arguably been one of the most controversial topics in the philosophy of biology. Philosophers and biologists alike have long debated over "correct" concepts of species and their ontological status. The traditional account took species as a category, class, or type instantiated by individual organisms. After the advent of evolutionary theory, the typological concept came under fire by those who identify species with a part of biological lineage (Ghiselin 1974; Hull 1976). They forcefully

Seattle, WA; 1-4 November 2018 -479-

argued that a species is not an abstract type but a concrete historical entity of which individual organisms are mereological bits. Although this individualist thesis became a de-facto standard in the philosophy of biology in the last century, some have complained its lack of explanatory power and called for a revival of a type or natural-kind based concept of biological species (Boyd 1999).

To this debate between individualists and typologists, this paper introduces yet another thesis according to which species taxa are models of scientific theory. Model is a notoriously equivocal concept, but in this paper it is understood as a set-theoretic entity that makes sentences of a given theory true or false. This implies that biological species are mathematical, rather than biological or even tangible, entities. To work out this claim I begin Section 2 with a reconstruction of various (e.g., phenetic, cladistic, biological etc.) species concepts in terms of formal models that licence characteristic sets of inferences. The model-theoretic rendering illustrates logical connections among different species concepts and provides a platform to evaluate them as a problem of *model selection*. Section 3 then expounds on philosophical implications of the model-theoretic interpretation. Identifying species with models entails that the organism-species relationship is not instantial or mereological, but rather representational; i.e., species as models *represent* individual organisms. This opens the possibility of applying general philosophical discussions on scientific representation and realism to vexed questions concerning the epistemic and ontological status of biological species. Through these arguments this paper puts the species problem under broader contexts of model selection, scientific representation, and scientific realism, depicting it as a special case of the generic question as to how science investigates the world.

2 Species as models

This section fleshes out the main claim of this paper by reconstructing various species concepts as set-theoretic models. The central idea is that species concepts specify theories that underpin biological inferences and descriptions, and species are models that satisfy such theories.

2.1 Typological species concepts

The traditional typological view defines species by its essence, or necessary and sufficient conditions or traits. This finds a straightforward expression as a biconditional form $\forall x(Sx \leftrightarrow T_1x \wedge T_2x \wedge \cdots)$. The extension of species S that satisfies this formula then is the intersection $\bigcap_i \mathbf{T}_i$ (see Figure 1(a)).

Though crude as it is, the biconditional formulation allows certain inferences from traits to species and vice versa. It is this kind of logical reasoning that has enabled, for example, the famous French zoologist George Cuvier to reconstruct the anatomy of a whole organism from just a single piece of bone. As is well known, however, such inferences have very restricted validity, because in most cases it is impossible to find a definite set of phenotypic or genetic characteristics that exclusively defines a given species. Evolution implies species boundaries to be necessarily "fuzzy," which undermines simple biconditional forms. The typological species concept has thus been criticized for its lack of expression ability: a simple algebra of trait-sets cannot capture the nuanced reality of biological species.

(a) Typological Sets (b) Cluster Phenotypic space (c) Cladistic Posets (d) Relational Graphs

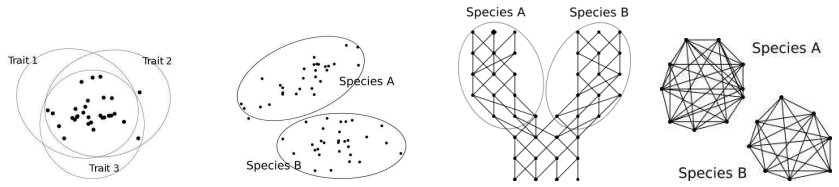


Figure 1: Illustrations of models of various species concepts, with corresponding formal setups. In each model dots/nodes represent individuals. See text for explanation.

2.2 Cluster species concepts

The cluster species concepts avoid this difficulty by defining a species as a group or cluster of similar organisms that do not necessarily share a common set of traits. The question then is how to define similarity. Its earliest variant, the phenetic species concept, represents organisms in a multi-dimensional space each axis of which defines a recorded trait (Sokal and Sneath 1963). Phenotypic similarity is then measured by the euclidean distance between two points/organisms, and a chunk or cluster of organisms in this euclidean space is identified as a species (Figure 1(b)). The choice of euclidean distance is not obligatory. One could, for example, measure similarity by the cosine between two points in the normalized phenotypic space, in which case the similarity amounts to correlation, with a species being identified as a correlated cluster or more generally a *probability distribution* over the phenotypic space (Boyd 1999).

The phenotypic space with a certain metric or probability distribution is certainly a much richer machinery than overlapping sets and allows for more nuanced expressions and inferences. The sophisticated theoretical background (euclidean geometry or

probability theory) enables one to measure the similarity among organisms and to make a trait-species inference in the absence of necessarily or sufficient criteria. To what extent such clustering and inference reflect objective species boundaries, however, was disputed, for the similarity calculation depends much on which phenotypic characters are taken into account. It should also be noted that, like the typological concept, the cluster concepts are purely static and lack a means to express the evolutionary past, the point often criticized by more historical approaches to species.

2.3 Cladistic species concepts

The cladistic species concepts focus on evolutionary history and define species solely in terms of phylogenetic relationships, as a "branch" (monophyletic group) in the evolutionary tree (Hennig 1966). Since ancestral relationship is antisymmetric and transitive, phylogeny forms a (strict) *partially ordered set* or *poset* $(\Omega, \prec)$, with Ω corresponding to a set of organism and $\prec$ meaning "is an ancestor of." A cladistic species is then defined as descendants from some founder organism(s) ω_f:

$$\{\omega \in \Omega : \omega_f \prec \omega\}. \tag{1}$$

An obvious advantage of the cladistic concepts is that it is faithful to the fact of evolution, and for this reason it has been most well received by biologists and philosophers alike. It is not, however, without flaws. For one, although the requirement of monophyly specifies a necessarily condition, it is silent as to how big a branch must be to qualify as a species (for even a small family can satisfy (1)), and so far no satisfactory sufficient condition was given (Velasco 2008). The monophyly requirement has also been

criticized to be too strong, for it would count birds as reptiles because the smallest monophylic group including lizards, snakes, and crocodiles also includes birds. That is, the cladistic species concepts make paraphyletic groups like reptilia *meaningless* (*Sensu* Narens 2007), which strikes some to be too high a price to pay.

2.4 Relational species concepts

Another popular approach is to define a species as a group of individuals in a certain relationship to each other. The biological species concept, for instance, defines species as "groups of interbreeding populations that are reproductively isolated form other such groups (Mayr 1942)" so that the required relationship here is mutual crossability. Other variants focus on reproductive competition (Ghiselin 1974) or organisms' capacity to recognize each other as a possible mate (Paterson 1985). All these proposals try to reduce species into mutual relationships (interbreeding, competition, recognition, etc.) between a pair of organisms. If we represent such relationships by an edge between nodes/organisms, a relational species can be defined as an isolated complete subgraph or *clique* in an undirected graph, that is, a group of nodes in which every two distinct nodes are connected but none is connected to outside (Figure 1(d)). Relational species thus find their model in graph theory, where edges represent the relation in question.

A common criticism of relational species concepts is that the focal relationship such as crossability sometimes fails to induce isolated cliques because some organisms at a species boundary can often mate with organisms that are thought to belong another species (e.g. ring species). Moreover, the biological species concept has been criticized to imply every asexually reproducing organism forms a distinct species (for any singleton

node is complete). These criticisms suggest that the real biological network is so "messy" that just a single relationship cannot divide it into distinct cliques in a non-trivial way.

2.5 "Combo" solutions

The model-theoretic rendering makes explicit what each species concept can and cannot meaningfully say about the biological world. Given that most of the criticisms we have seen concern the "cannot say" part, one way to deal with these difficulties is to combine different theories to obtain more complex definitions of species.

For instance, one may combine the cluster and cladistic species concepts and define a species as a *lineage that shares the same or similar phenotypic distribution*:

$$\{\omega \in \Omega : \omega_f \prec \omega \wedge \theta(\omega_f) = \theta(\omega)\} \tag{2}$$

where $\theta : \Omega \to \mathbb{R}^n$ assigns distribution parameters to each organism $\omega \in \Omega$.[1] On this definition one may meaningfully define paraphyletic species and distinguish birds from other reptiles on the basis of the difference in their phenotypic or genetic profiles. It can also account for anagenesis (speciation without branching) and continuity of species between a cladogenesis (splitting event).

If one replaces θ in (2) with a different function $\nu : \Omega \to N$ that maps organisms $\omega \in \Omega$ to their *niche* $\nu(\omega) \in N$, it becomes the *ecological species concept* which defines a species as "a lineage ... which occupies an adaptive zone minimally different from that of

[1]For non-parametric cases, we can set $\theta : \Omega \to \mathbb{R}^\infty$ and modify the definition as $\{\omega \in \Omega : \omega_f \prec \omega \wedge D(\theta(\omega_f), \theta(\omega)) < k\}$ where $D(\bullet)$ is a divergence measure (such as the Kullback-Leibler divergence) and k is a constant.

any other lineage in its range (Van Valen 1976, 233)."

Yet another combination is that of the cladistic and biological species concepts, which would define a species as a maximum monophylic lineage that can mutually interbreed, so that

$$\{\omega_x, \omega_y \in \Omega : \omega_f \prec \omega_x \wedge \omega_f \prec \omega_y \wedge \omega_x \sim \omega_y\} \tag{3}$$

where $\sim$ stands for crossability.[2] This will make up for the lack of a sufficient condition in the cladistic species concept, and accord well with the so-called *evolutionary species concept* which emphasizes the unique "evolutionary tendencies and historical fate" of each species (Wiley 1978, 17). It should be noted that this could also avoid the problem of ring species because two crossable organisms may not necessary share the same ancestor.

2.6 The scientific species problem as a problem of theory choice

The above discussion shows that (i) major species concepts can be defined as models of formal theories, and that (ii) more complex concepts can be obtained by combining basic ones. The model-theoretic approach characterizes each species concept with the formal apparatus it assumes, which in turn determines its expressive power or what can meaningfully be stated about organisms and/or their history (Narens 2007). In general, a richer theoretical apparatus allows for more nuanced expressions, which makes it less liable to counterexamples. This is illustrated in the progression from the typological to

[2]As in the case of the biological species concept, the crossability here must take into account the existence of two sexes.

cluster and then to cluster-cladistic concepts, where in each step the species concept acquires the ability to deal with fuzzy boundaries and evolutionary history, respectively.

It does not necessarily follow, however, that a richer concept is always desirable, because it tends to have a greater degree of freedom and requires more data in actual application. While only phylogenetic information suffices to demarcate cladistic species, the cluster-cladistic concept also requires phenotypic or ecological information, which in many cases may not be available. A stronger semantic power thus comes with a higher epistemic cost, as is often emphasized by pheneticists or cladists in their respective advocacy of the phenotypic cluster and cladistic species concepts.

This suggests that the competition among various species concepts should be understood as a problem of model selection, where different models are evaluated on the basis of their explanatory or descriptive power versus parsimony or operationality (Sober 2008). Indeed, most disputes among advocates of different species concepts arise from their differential emphasis on what aspects of the biological world a disable species concept needs and needs not take into account (Ereshefsky 2001), but the difficulty is that these emphases are often implicit and incommensurable. Although the model-theoretic approach does not arbitrate these debates, it provides a common formal framework that makes explicit the explanatory power and operationality of species concepts and facilitates evaluation of their respective advantage.

3 Philosophical implications

3.1 Species are models

Upon the model-theoretic reconstruction of various species concepts, we now turn to the philosophical thesis that species taxa should be construed as models proposed above, i.e., as set-theoretic entities. To proceed, let me first begin with an analogy from classical mechanics. Classical mechanics is a theory about Newtonian particles, which are customary defined as volumeless points or vectors in a three-dimensional Euclidean space. Newton's celebrated laws like $\mathbf{F} = m\mathbf{a}$ describe temporal evolution of a system composed of such "particles." This system is to be distinguished from any actual physical systems, say the solar system, for one thing, no concrete bodies are volumeless, nor do they indefinitely continue rectilinear motion as prescribed by Newton's first law. Newton's theory, or any other physical theories for that matter, is a description of idealized and abstracted models and not of actual phenomena (Cartwright 1983). That is, models of classical mechanics — which make its laws and statements true — are not concrete, physical entities, but rather abstract mathematical objects that can be constructed within set theory (McKinsey et al. 1953).

The role of models in science has been emphasized by the so-called semantic or model-based view of scientific theories (e.g. van Fraassen 1980; Suppe 1989).[3] In the traditional, logical-positivist view, a scientific theory was supposed to directly describe

[3]This label ("the semantic view") has been used to describe different, and logically independent, theses. In particular, while some philosophers (e.g. Suppes 2002) take a scientific theory as a *description* of models, others *identify* it with a set of models (van Fraassen 1980). In this paper I adopt the former thesis without committing to the latter.

observed data. This has set for positivists the difficult task of reducing theoretical concepts that seemingly lack direct empirical contents to observation vocabulary by way of *bridge laws* or *partial interpretations*. To avoid this difficulty, proponents of the model-based view take a model, rather than observation, as the primary descriptive target of a scientific theory. In this view, a theory specifies an abstract model that idealizes and extracts just salient factors, and only indirectly relates to actual phenomena via such an model.

I submit that the species problem is a variant of the positivist conundrum. Species is a highly theoretical concept, and various proposal of "species concepts" in the past can be understood as attempts to build bridge laws for reducing it to a set of observational or operational criteria. To date more than a dozen of different concepts have been proposed[4], with no general consensus — each has its own strength, but also weakness and exceptions when applied to the rich and heterogeneous biological world. The assumption has been that a species concept must be a faithful description of *actual* biological features or phenomena. But what if this assumption is untenable, or at least unreasonable? The model-based view has been quite popular among philosophers of biology (e.g. Beatty 1981; Lloyd 1988). If we adopt this view and construe evolutionary theory as describing models, then species too must be defined accordingly, i.e., as (a part of) abstract models that satisfy descriptions and/or inferences of the corresponding theory.

What, then, are theories about species? Without claiming to be exhaustive, this paper adopts Suppes's (2002) thesis that a scientific theory must be defined as a set-theoretical predicate. The foremost advantage of this approach is that it enables one

[4]Mayden (1997), for example, counts at least 22 concepts of species.

to easily harness a theory with mathematical apparatus necessary for sophisticated reasoning. As discussed above, contemporary studies on species rely heavily on quantitative methods to calculate similarity or reconstruct a phylogenetic tree from phenotypic or genetic data. Given that such mathematical reasoning requires matching formal models of calculus or probability theory, the straightforward way to define a species is to build it upon these mathematical backgrounds as an extension of these formal models. Section 2 is a preliminary sketch of applying this Suppesian program to various species concepts. If this attempt turns out to be successful, biological species are to be understood as parts of set-theoretic structures, just like Newtonian particles. That is, they are mathematical and abstract constructs, rather than physical or biological entities.[5]

The purpose of the set-theoretic exposition is not just to accommodate quantitative reasoning. Even with less quantitative cases like the biological species concept, it makes implicit assumptions explicit and suggests a way to deal with counterexamples. The problem of ring species, for example, arises from a conflict between the presumption that each biological species must be isolated and the fact that crossability is not necessarily transitive and thus fails to induce equivalence classes. One possible response to this charge then would be to weaken the former assumption and redefine a species just as a (not necessarily isolated) clique in the reproductive network. Clarification of theoretical assumptions helps us to assess other species concepts as well. For example, the phenetic species concept is often claimed to be "theory-free" in that it does not depend on any evolutionary hypothesis. But as we have seen in Sec. 2.2, the calculation of phenotypic

[5]Hence the present thesis should not be confused with the view that species are sets or collections of *organisms* (Kitcher 1984), which, after all, are concrete biological entities.

similarity presupposes a phenotypic space equipped with a particular (e.g., euclidean) metric, which is a fairly strong theoretical assumption. Also, cladists often stress the simplicity and purity of their monophylic species definition that only considers phylogenetic relationships. But in order to make use of likelihood methods to infer such relationships, as is common in practice, a simple poset is not enough: one also needs to assume some genetic or phenotypic distribution, and then there is no in-principle reason to exclude non-monophylic taxa from the definition of species (as (2) in Sec. 2.5).

The final but not least merit of the set-theoretic approach is its flexibility: it allows for a construction of a new species concept by combining existing ones (Sec. 2.5) or adding new theoretical assumptions. For instance, it is common in experimental biology to characterize a species by shared developmental or causal mechanisms: developmental biologists often talk about "the development of the chicken" and medical doctors rely on causal extrapolation when they prescribe a clinically-tested drug for their patient. Such a "causal species" may be defined by isomorphic *causal models*, which combine a probabilistic distribution and a causal graph over variables. Hence the discussion in Section 2 covers just a few samples that can be constructed within this general framework. This does not of course mean that every possible species concept can and must be formalized, but does suggest the potential of the set-theoretic approach to accommodate the use of existing species concepts and to develop novel ones.

3.2 Philosophical implications

Identifying species with theoretical models sheds new light on some vexed philosophical issues, one amongst which concerns how individual organisms are related to species taxa.

Philosophers have long debated whether the organism-species relationship is instantial (organisms are particular *instances* of a species *qua* class), membership (they are *members* of a species *qua* set; Kitcher 1984), or mereological (they are *parts* of a species *qua* genealogical entity; Ghiselin 1997). The model-theoretic approach suggests an alternative account, according to which a species *represents* (a group of) individual organisms. Just as the Rutherford-Bohr model represents the microscopic structure of atoms, models proposed in Section 2 represent biological populations: for example, nodes and edges consisting of the biological species model in Figure 1(d) respectively represent organisms and crossability. Representation captures our intuitive notion that a model and its target phenomenon share salient static or dynamic features up to a certain precision. Given that said, it must be admitted that the criteria and nature of scientific representation are diversified and still open questions (Frigg and Nguyen 2016). Hence calling the species-organism relationship representational does not necessarily demystify it, but at least implies that the problem is not endemic to evolutionary theory: it is rather a version of a broader philosophical issue as to how the use of scientific models help us understanding the world. This means that the arsenal of this rich philosophical literature can and should be consulted to elucidate the nature of the species-organism relationship. Another, more immediate implication is that the membership and mereological accounts must be both abandoned, for whatever the relationship between a model and phenomena turns out to be, the latter must certainly not be a member or part of the former.

Neither is representation identity or instantiation. Ideal gas is not identical to any actual gas, but only approximates thermodynamic characteristics of some. Hence strictly speaking it has no instantiation, but this does not detract its epistemic validity. Likewise

species concepts, as specifications of ideal models, need not directly apply to actual populations. No wild population big enough to qualify as a species would strictly satisfy the requirement of the biological species concept, because actual mating chance is often hindered by physiological, geographical, and other contingencies. In the same vein, a phenetic or genetic cluster is expected to have outliers when applied to a real population. However, the presence of such exceptions should not immediately invalidate the corresponding species concepts, because the value of a species concept consists less in its universal validity than its epistemic serviceability for inferences and explanations of evolutionary or biological phenomena. These two criteria often conflict: Cartwright (1983) even argues that explanatory theories necessarily distort the reality by idealizing the situation and extracting only relevant features, so that properly speaking they are "lies" by design. Cartwright's examples are physics and economics, but her idea also applies to the present context. The primary function of a species concept is to explain biological phenomena rather than to save them, so that a few discrepancies should not be taken as a falsification.

The conflict between exceptionlessness versus explanatory power also underlies the realism-nominalism debate over species. The proponents of the nominalistic thesis who claim a species to be nothing but a totality of individual organisms have motivated their view by criticizing the realist interpretation of species-as-class for its commitment to the typological thinking and failure to deal with the evident heterogeneity of biological phenomena (e.g. Ghiselin 1997). On the other hand, those who attach weight on the role of species concept in induction and explanation have upheld a realist position and treated species as natural kinds (Boyd 1999). The present thesis offers a third alternative, recognizing the explanatory role of species concept without committing to

Seattle, WA; 1-4 November 2018 -493-

the ontologically heavy assumption of natural kinds. As we have seen in Section 2, species as models licence particular sets of inferences. The cluster and typological species/models underpin an expectation that physiological or genetic features found in, say, laboratory animals would also be shared by other individuals of the same species, while the evolutionary species concept explains the reason of such intra-specific similarities. These explanations are effectuated by the same model representing numerically distinct individuals or phenomena to be explained. Note that this procedure no more presupposes the existence of the model as an independent, real entity, than do explanations based on, say, ideal gas. Indeed, explanations may be based on fictional models, as is the case with the Ising model in statistical mechanics.

This does not of course mean that models *must be* fictions, or that species do not exist. Recent advocates of scientific realism argue that successful scientific models capture some, especially structural, aspect of reality (Ladyman 2016). Given its affinity to the model-based view of scientific theories, species realists may well apply this line of reasoning to the present context, taking the set-theoretic structures as discussed in Section 2 as representing the reality or "essential feature" of biological species. Whether and to what extent such an argument carry over, however, remain to be examined by a further study.

4 Conclusion

The past debates over biological species have been based on the assumption that species concepts must describe actual biological phenomena, the strict adherence to which tends to rule out all but cladistic species as typological or inexact. The present paper

challenged this assumption and argued that the primary referent of a species concept is a (set-theoretic) model that licences a certain set of inferences specified by the concept. The model-theoretic rendering articulates explanatory power and theoretical assumptions of each species concept and illuminates logical relationships among them. Once species are specified as models, the long-standing competition among different species concepts reduces to a common problem of model selection. This suggests that evaluation of relative merits and demerits of species concepts must be based more on their explanatory power than on exceptionlessness.

On the philosophical side, the shift in the ontological status of species means that the organism-species relationship is not that of instantiation, membership, or mereology, but rather representation. The vexed issue that has troubled philosophers for decades, therefore, boils down to the broader problem as to how and why scientific models can be used to represent and explain the world. This suggests the possibility to apply the rich literature on scientific representation and realism to elucidate the epistemological and ontological nature of biological species.

In sum, the take home message of the present paper is that the species problem is not endemic to biology or evolutionary theory, but rather is a variant of general scientific and philosophical issues of model selection, scientific representation, and realism. The purpose of this paper was just to establish such a parallelism: determining its philosophical implications on specific debates such as realism or pluralism concerning biological species will be a task for future studies.

References

Beatty, John. 1981. "What's Wrong with the Received View of Evolutionary Theory?." *PSA 1980* 2: 397–426.

Boyd, Richard N. 1999. "Homeostasis, species, and higher taxa." In *Species: New Interdisciplinary Essays.* ed. Robert A Wilson, 141–158, Cambridge, MA: MIT Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Ereshefsky, Marc. 2001. *The Poverty of the Linnaean Hierarchy.* Cambridge: Cambridge University Press.

van Fraassen, Bas C. 1980. *The Scientific Image.* Oxford: Oxford University Press.

Frigg, Roman, and James Nguyen. 2016. "Scientific Representation." In *The Stanford Encyclopedia of Philosophy.* ed. Edward N Zalta, Metaphysics Research Lab, Stanford University.

Ghiselin, Michael T. 1974. "A Radical Solution to the Species Problem." *Society of Systematic Biologists* 23: 536–544.

——— 1997. *Metaphysics and the Origin of Species.* Albany, NY: State University of New York Press.

Hennig, Willi. 1966. *Phylogenetic Systematics.* Urbana, IL: University of Illinois Press.

Hull, David L. 1976. "Are species really individuals?" *Systematic Zoology* 25: 174–191.

Kitcher, Philip. 1984. "Species." *Philosophy of Science* 51: 308–333.

Ladyman, James. 2016. "Structural Realism." In *The Stanford Encyclopedia of Philosophy.* ed. Edward N Zalta, Metaphysics Research Lab, Stanford University.

Lloyd, Elisabeth A. 1988. *The Structure and Confirmation of Evolutionary Theory.* Princeton, NJ: Princeton University Press.

Mayden, R L. 1997. "A hierarchy of species concepts: the denouement in the saga of the species problem." In *Species The Units of Biodiversity.* ed. M F Claridge, H A Dawah, and M R Wilson, 381–424, London: Chapman & Hall.

Mayr, Ernst. 1942. *Systematics and origin of species.* New York, NY: Columbia University Press.

McKinsey, John C C, Patrick Suppes, and A C Sugar. 1953. "Axiomatic Foundations of Classical Particle Mechanics." *Journal of Rational Mechanics and Analysis* 2: 253–272.

Narens, Louis. 2007. *Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science.* Mahwah, NJ: Lawrence Erlbaum Associates.

Paterson, Hugh E H. 1985. "The Recognition Concept of Species." In *Species and Speciation.* ed. E. S. Vrba, 21–29, Pretoria.

Sober, Elliott. 2008. *Evidence and Evolution.* Cambridge: Cambridge University Press.

Sokal, Robert R, and Peter H A Sneath. 1963. *Principles of Numerical Taxonomy.* San Francisco, CA: W. H. Freeman and Co.

Suppe, Frederick. 1989. *The Semantic Conception of Theories and Scientific Realism.*: University of Illinois Press.

Suppes, Patrick. 2002. *Representation and Invariance of Scientific Structures.* Stanford, CA: CSLI Publication.

Van Valen, Leigh. 1976. "Ecological Species, Multispecies, and Oaks." *Taxon* 25: 233–239.

Velasco, Joel D. 2008. "The internodal species concept: a response to 'The tree, the network, and the species'." *Biological Journal of Linnean Society* 93: 865–869.

Wiley, Edward O. 1978. "The Evolutionary Species Concept Reconsidered." *Systematic Biology* 27: 17–26.

Historical Inductions Meet the Material Theory

by Elay Shech

Oct. 2018
(Pre-conference version)

Forthcoming in *Philosophy of Science*

Acknowledgements: I am indebted to John Norton and Moti Mizrahi for extremely valuable discussion and comments on earlier drafts of this paper. Thank you also to helpful conversation with the audience at the Auburn University Philosophical Society in the Spring of 2018 and participants in Gila Sher's *Truth and Scientific Change* reading group in the Fall of 2017 at the Sidney M. Edelstein Center for History and Philosophy of Science, Technology and Medicine at the Hebrew University of Jerusalem.

Abstract: Historical inductions, viz., the pessimistic meta-induction and the problem of unconceived alternatives, are critically analyzed via John D. Norton's material theory of induction and subsequently rejected as non-cogent arguments. It is suggested that the material theory is amenable to a local version of the pessimistic meta-induction, e.g., in the context of some medical studies.

1. Introduction

My goal is to contribute to a growing literature that is critical of historical inductions such as the pessimistic (meta-)induction (PMI) argument (Poincaré 1952, 160; Putnam 1978, 25; Laudan 1981) and the problem of unconceived alternatives (Stanford 2001, 2006) against scientific realism, concentrating mostly on the former. The PMI can be construed in different ways (Mizrahi 2015, Wray 2015), viz., as a deductive *reductio ad absurdum* (e.g., Psillos 1996, 1999), a counterexample to the no miracles argument and inference to best explanation argument for scientific realism (e.g., Saatsi 2005, Laudan 1981), or, usually, as an inductive argument (e.g., Poincaré 1952, Putnam 1978, Laudan 1981, Rescher 1987). In the following I will argue against the inductive version of PMI—or any construal of the PMI that makes use of historical induction—using John D. Norton's material theory of induction (Norton 2003, Manuscript). The upshot is that one ought to be critical of historical inductions that seem to fit the general form or pattern of a good inductive argument, but may in fact lack inductive warrant and force. Various critiques have been put against the PMI (e.g., Lange 2002, Lewis 2001, Mizrahi 2013), along with some defenses (e.g., Saatsi 2005). In Section 2 I will present the PMI and briefly discuss some criticism in order to place my own analysis in broader context. Section 3 presents the material theory of induction and argues that it dissolves the PMI, while Section 4 extends such claims to the more recent problem of unconceived alternatives. In Section 5 I note that the material theory of induction does leave room for a local version of the PMI, which holds in some

limited domain, such as in relation to certain medical studies (Ruhmkorff 2014). I end in Section 6 with a short conclusion.

2. The (Inductive) Pessimistic (Meta-)Induction

The modern formulation of the PMI is usually attributed to Laudan (1981) who argued that having genuinely referential theoretical and observational terms, or being approximately true, is neither necessary nor sufficient for a theory being explanatory and predictively successful. More generally, Anjan Chakravartty characterizes the argument as follows:

> [PMI can] be described as a two-step worry. First, there is an assertion to the effect that the history of science contains an impressive graveyard of theories that were previously believed [to be true], but subsequently judged to be false . . . Second, there is an induction on the basis of this assertion, whose conclusion is that current theories are likely future occupants of the same graveyard. (Chakravartty 2008, 152)[1]

The PMI then may take the following form:

> **[Inductive Generalization PMI]**
>
> P(i) Past theory 1 was successful but not genuinely referential or approximately true.
> P(ii) Past theory 2 was successful but not genuinely referential or approximately true.
> …
> C) Therefore, current (and perhaps future) theories are successful but (by induction) probably not genuinely referential or approximately true.

Laudan (1981) suggests that the history of science contains a graveyard of theories that were previously believed to be approximately true and genuinely referential, but that subsequently were judged to be false and not to refer. Estimations of the number of such superseded theories have been debated (e.g., Lewis 2001, Wray 2013) and recently Mizrahi (2016) presents evidence that challenges the "history of science as a graveyard of theories" claim. Others voice concerns regarding the period of history of science used in order to extract historical evidence (e.g., Lange 2002, Fahrbach 2011) or the proper unit of analysis, i.e., theories vs. theoretical entity (e.g., Lange 2002, Magnus and Callender 2004). Similarly, Park (2011, 83) and Mizrahi (2013, 3220-3222) have argued that the PMI is fallacious due to cherry-picking data, biased statistics, and non-random sampling.

My own criticism of the inductive PMI comes from a different avenue. I will assume that the anti-realist does have randomly sampled historical evidence from the correct period of history and with the proper unit of analysis (whatever those

[1] cf. Wray (2015, 61).

may be) that is not biased or cherry-picked. Still, on the material theory of induction the PMI will not be a cogent argument. In other words, I aim to identify what I take to be a more fundamental (although not categorically different) problem with the PMI.

3. PMI Meets the Material Theory

3.1 The Material Theory of Induction in a Nutshell

Consider the following formally identical inductive inferences (Norton 2003, 649):

> P1) Some samples of the element bismuth melt at 271 degrees C.
> C1) Therefore, all samples of the element bismuth melt at 271 degrees C.
>
> P2) Some samples of wax melt at 91 degrees C.
> C2) Therefore, all samples of wax melt at 91 degrees C.

What makes the first argument an inductively strong and cogent argument while the second a weak and non-cogent inductive argument? Norton (2003, Manuscript) has argued that formal theories of induction, which provide universal schemas that are meant to identify the inductions that are licit and those that are not, stand against an insurmountable difficulty when facing such a question.[2] Instead, he offers a material account of induction:

> In a material theory, the admissibility of an induction is ultimately traced back to a matter of fact, not to a universal schema. We are licensed to infer from the melting point of some samples of an element to the melting point of all samples by a fact about elements: their samples are generally uniform in their physical properties. … *All inductions ultimately derive their licenses from facts pertinent to the matter of the induction*. (Norton 2003, 650; original emphasis)

Norton calls the local facts that power inductive inferences "material postulates." Material postulates themselves are supported by other instances of induction that are licensed by different material postulates.

3.2 Material Analysis of PMI

Many of the criticism of the inductive PMI discussed above amount to the claim that the universal schema used by the likes of Laudan (1981), namely, (P3) Some A's are B's, (C3) Therefore, all A's are B's, does not apply in the case of the PMI because various criteria needed to implement the scheme, e.g., random sampling, correct historical period, proper unit of analysis, have not been met. What I wish to do here

[2] I will not defend Norton's theory or claims here. He dedicates an entire book to the matter in Norton (Manuscript).

Seattle, WA; 1-4 November 2018 -501-

is conduct a material analysis of the PMI. Considering the above presentation of the PMI in its [Inductive Generalization PMI] form we may ask, what powers the inductive inference, i.e., what material postulate licenses the pessimistic conclusion?

In context of the two inductive arguments considered in Section 3.1, we note that there is no material postulate that licenses the inductive inference in the case of wax (P2 too C2) but there is one in the case of bismuth (P1 to C1): Generally, chemical elements are uniform in their physical properties. By analogy, the presumption of the meta-induction is that each historical case study looked at is an instance of the same thing, a discovery of induction in science. If we are to perform the meta-induction then there needs to be something in the background facts that unifies all such inductions, just like the fact chemical elements are generally uniform in their physical properties warrants the inductive inference regarding the melting point of bismuth. Let us consider several options.

First, perhaps the material fact is that most scientists use a common rule or method in constructing or discovering successful theories, something along the lines of Mill's methods of experimental inquiry in his *System of Logic* (1872, Book III, Ch. 7). If so, the properties of the rule would be used to authorize the induction. Is there such a rule, or perhaps, some common scientific method? A glance at the history of science suggests that this is unlikely. Newton's deduction from the phenomena, is very different from Darwin's inference to best explanation, which in turn differs radically from Einstein's thought experiments with lights beams, trains, and elevators.[3] More generally, there seems to be a consensus among historians and philosophers of science that something like "the scientific method" is really more of an umbrella term for very different methods used by scientists to construct and discover theories. After all, novel problems necessitate novels solutions, and the commonality that does arise in different cases, say, attempts to minimize error or to be objective, is not the kind of commonality that we seek in powering the PMI and drawing the pessimistic conclusion. For instance, in his book *Styles of Knowing: A New History of Science from Ancient Times to the Present*, Chungling Kwa (2011) argues that there is no single, fundamental method used in science: "there is not just one form of Western scientific rationality; there are at least six." The framework of six "styles of knowing," includes the deductive, the experimental, the hypothetical-analogical, the taxonomic, the statistical, and the evolutionary style, and is based on Alistair Crombie's (1994) three-volume work *Styles of Scientific Thinking*. Similar, Ian Hacking (also taking lead from Crombie's work) has argued that there are distinct "styles of reasoning" used in science, such as the postulational style, the style of experimental exploration, the style of hypothetical construction of models by analogy, the taxonomic style, the statistical style, the historical derivation of genetic development, and the laboratory style (Hacking 1992). This further

[3] In fact, see Norton (Manuscript, Ch. 8-9) who argues that even in historical cases where the *same* principle is applied by scientists, viz., inference to best explanation, "at best we can find loose similarities that the canonical examples of inference to best explanation share," so that no common rule of the kind needed to power the PMI can be found (Ch. 8, p. 1).

corroborates the idea that scientific methods used for theory construction and discovery, as well as for scientific explanation, are very diverse.

More generally, scientific theories are not kind of things that portray the type of uniformity needed to license inductive inferences on Norton's material theory. Albeit in a different context, a similar point is nicely made by Mizrahi (2013, 3218):

> A uniform—as opposed to diverse—sample might be a sample of, say, copper rods. From a sample of just a few copper rods that are tested for electrical conductivity, it is reasonable to conclude that all copper rods conduct electricity because, if you have seen one or two copper rods, you have seen them all (given their uniform atomic structure). Scientific theories, however, are not as uniform as copper rods. The point, then, is that any sample of theories is not going to be uniform in a way that is required for a "seen one, seen them all" inductive generalization.

Similarly, and second, perhaps there are some facts about investigating scientist themselves, how they work, and/or the problems situations that they work in, which can unify the historical evidence in a way that provides us with the inductive warrant we seek. Maybe such facts will include something about the psychology of scientists: their fastidiousness and fear of error, their facility at jumping to conclusions, or perhaps their curiosity, logic, creativity, skepticism, etc. However, in a similar manner to the search for a common rule used in constructing successful theories, the history of science furnishes us with scientists that are heterogeneous enough in their psychological traits, and work in such varied contexts, so as not to provide us with any was to unify the various historical cases in a way pertinent to licensing the pessimistic inference of the PMI.

Third, perhaps we can circumvent looking to a common rule of constructing or discovering theories, or searching for common traits among scientists, by noting that the follow candidate material postulate would power the PMI:

> MP-PMI: Generally, successful theories are not genuinely referential and/or approximately true.

But how would we establish MP-PMI? One option is to appeal to the PMI itself, but this would either be circular or else push us to look for another material postulate. Another option is just to grant the MP-PMI as a reasonable assumption. Perhaps anti-realists or instrumentalists would think that this is a sensible starting point, but their target realist opponent would surely reject such an assumption as question begging. Last, perchance there is some fact about explanatory and/or predictively successful theories that renders them, generally, not genuinely referential and/or approximately true? Possibly part of the essence of successful theories is to misrepresent the world? To me this seems highly unlikely and at odds with any levelheaded intuition but, in any case, if we could argue that successful theories are essentially inaccurate then we would not need the PMI in the first place!

Fourth, we may want to construe the PMI in its inductive generalization form as a kind of abductive argument with the following type of material postulate:[4]

> **[Inductive Generalization PMI – Abductive version]**
>
> P(i): The success of past theory 1 (constructed using method m) is not best explained by its truth.
> P(ii): The success of past theory 2 (constructed using method m) is not best explained by its truth.
> …
> MP: Scientific theories constructed using method m are generally uniform with respect to what best explains their predictive success.
> C: The success of our best current (and perhaps futures) theories (constructed using method m) are not best explained by their truth.

Stating the PMI as above has the merit of directly engaging with the "no miracles argument" for scientific realism, namely:

> That terms in mature scientific theories typically refer [to things in the world] …, that theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when it occurs in different theories—these statements are viewed by the scientific realist not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate scientific description of science and its relations to its objects. (Putnam 1975, 73)

But worries abound. First, the realist may very well deny P(i), P(ii), etc., and argue that the success of past theories is best explained by their truth but that, as it turns out, either the best explanation did not hold in this case or else there is some sense in which past theories, insofar as they were successful, were approximately true or on the road to truth. Second, construing the argument as an abduction opens up a Pandora's box of problems associated with the notion of explanation: What is explanation? Are there accounts of explanation where success is best explained by truth and ones in which it isn't and, if so, which account of explanation is relevant in this context? And so on.

Third, the cogency of the argument depends on the idea that all theories appealed to were constructed with some method m, but we already judged that there is no one method that is relevant to constructing scientific theories. Perhaps phenomenological models are good candidates for the type of things that can provide empirical success but are not generally approximately true.[5] Thus, at best, the above argument can power a kind of local PMI: Successful theories constructed

[4] Thanks to Tim Sundell for suggest this line of thought.
[5] Phenomenological models are, generally, not considered explanatory.

by method m are not approximately true. We'll consider one such case in more detail in Section 5.

In short, on the material theory of induction inductive arguments are powered by facts, by material postulates, but in the context of the PMI it seems unlikely that any such non-question begging postulates, which wouldn't render the PMI obsolete, can be found. This is so even if, say, the historical data was not cherry-picked, and the right unit of analysis and correct period of history were used. In other words, I'm equally skeptic of projects that attempt to block the pessimistic conclusion by, for example, taking a random sample of past scientific theories, e.g., Mizrahi (2016). In the following section I'll attempt to extend such claims to the problem of unconceived alternatives.

4. Extension to the Problem of Unconceived Alternatives

Recently, P. Kyle Stanford (2001, 2006) has developed what may be characterized as a new version of the PMI:

> … I propose the following New Induction over the History of Science: that we have, throughout the history of scientific inquiry and in virtually every field, repeatedly occupied an epistemic position in which we could conceive of only one or a few theories that were well-confirmed by the available evidence, while subsequent history of inquiry has routinely (if not invariably) revealed further, radically distinct alternatives as well-confirmed by the previously available evidence as those we were inclined to accept on the strength of that evidence. (Stanford 2001, S8-S9)

The problem of unconceinved alternatives as an argument against scientific realism has been criticized on various grounds (e.g., Chakravartty 2008, Devitt 2011, Mizrahi 2015), but my goal here is just to note that the discussion of Section 3 can be extended to this new version of the PMI, which can be construed as follows:

> P(i) In the past time of theory 1, theory 1 was successful but there were unconceived alternative theories that were as well supported by available evidence but with radically different ontology.
> P(ii) In the past time of theory 2, theory 2 was successful but there were unconceived alternative theories that were as well supported by available evidence but with radically different ontology.
> …
> C) Therefore, in present times, current theories are successful but (by induction) there probably are unconceived alternative theories that are as well supported by available evidence but with radically different ontology.

What we need for the material analysis is something like: Generally, successful theories are underdetermined by data due to possible unconceived alternative theories. In a similar fashion to the MP-PMI, we could look to some common rule used by scientists to conceive theories, or some common psychological traits among

scientist, that may ground the idea that successful theories are such that empirically adequate unconceived alternatives always exists. But for the same reasons discussed above, it seems unlikely that any such common rule or traits will be found. That said, perhaps cognitive facts about human scientists might support the inductive inference to the conclusion that we always miss some alternative theories, which in turn are consistent with the available evidence. What is attractive about this line of thought is that it does seem plausible that due to our cognitive limitations there are always "unconceived alternatives." However, mere cognitive limitations do not support the further conclusion that there are unconcieved alternative theories that are *consistent with available evidence*.

Alternatively, one may think that Stanford's new induction circumvents the material objection: modal reflections alone convince us that there are always unconceived alternative theories that can explain and predict empirical phenomena just as well or better than conceived theories. But how can we come to such a conclusion based on modal reflections alone? Isn't it conceivable if not possible that there would be a point in history with no unconceived alternatives and isn't conceivable if not possible that we are at such point in time in history? Moreover, it is unclear what to make of theory-independent modal claims (unless one has logical modality in mind, which isn't the case here). Certainly, we can talk about different physically possible worlds given a particular physical theory. For instance, various solutions to the Einstein field equations are taken to denote different possible universes according to relativity theory. But it isn't clear what is meant by different possible or alternative conceivable *theories* given no meta-theory as a constraint, so to speak.[6] In any case, if we know that unconceived alternative theories always exist based on modal reflections alone, then the historical induction is doing no work for us at all.

5. Room for a local, material pessimistic induction?

Although the material analysis given here may prompt us to be skeptical of historical inductions (insofar as one is moved by the material theory of induction), it can help us understand why *local* pessimistic inductions may be tenable. Specifically, I want to look at a recent discussion by Rumkorf (2014) who contends that meta-analyses in medicine such as Ioannidis' (2005a, 2005b), which show that a disconcertingly high percentage of prominent medical research findings are refuted by subsequent research, can be developed into a local pessimistic induction. Ioannidis (2005a, 2005b) is concerned with studies, denoted "M-studies," that satisfy the following criteria: "being highly cited, using contemporary research and statistical methods, and being among the first studies to investigate a question at issue" (Rumkorf 2014, 420). Rumkorf's (2014, 421) then uses the various conclusions of Ioannidis (2005a, 2005b) to generate a local PMI in the field of medicine (PMI-M):

[6] What would count as a (logically possible but physically) impossible theory in such a context?

> E1 41% of the associative or causal claims made by M-studies in the sample were inconsistent with the results of subsequent published studies either (1) because the later studies provided evidence against the existence of the association or effect; or (2) because the later studies provided evidence that the magnitude of the association or effect was significantly different.
>
> E2 Therefore, we can expect approximately 41% of the associative and causal claims made by M-studies to be inconsistent with subsequent published studies.

On Norton's theory we need to appeal to a material postulate to license the pessimistic inductive inference in the transitions from E1 to E2, but since we are now working in a limited domain without many heterogeneous examples as in the whole history of science, we may now find some significant commonality between the methods used in different M-studies that can act as licensing facts. What are the background facts that power the PMI-M? Here are some options extracted from Ioannidis's diagnosis of his meta-analysis and quoted in Ruhmkorf (2014, 219):

> Contributing factors include: bias in research (Ioannidis 2005b); non-randomized trials (Ioannidis 2005a); smaller rather than larger sample sizes in refuted studies (Ioannidis 2005a, 224); and publication and time-lag biases (whereby studies with highly significant and potentially aberrational positive results are overrepresented among published articles in major journals and are published more quickly than other articles) (Ioannidis 2005a, 224). Particularly intriguing is the idea that large-scale features of the structure of medical and biological inquiry contribute to the high contradiction rate. Having a number of distinct working groups looking at the same problem increases the chances that at least one of them will find something statistically significant, especially if they are looking at a wide array of possible relationships (Ioannidis 2005b, 697–698). The computational power and richness of data sets available to researchers increases the chance that some of them will be successful in achieving statistical significance, even when no real relationship exists (Ioannidis 2005b, 701).[7]

These various factors, insofar as they are common to most M-studies, are the type of background facts that warrant the pessimistic induction from a material point of view. One may worry of course that the pessimism associated with local PMI generalizes since, presumably, facts about biases and the like are facts about researchers in general, not just researchers in medical science in particular. But, although all scientific studies have to deal with challenges such bias, it may be the case that a particular local subfield, due to its specific nature and whatever social

[7] It should be noted that there are some problems with Ioannidis's (2005a, 2005b) methodology, as identified in Ruhmkorff (2014, 419-421), but they do not seem to be problematic enough to render the PMI-M not cogent.

norms are in place for collecting and disseminative evidence, is especially challenged in a way that can justify the pessimistic induction. The above suggests that this is indeed the case for M-studies.

To end, Ruhmkorff (2014) argues against global PMI on independent grounds (namely, he argues that the PMI commits a statistical error previously unmentioned in the literature and is self-undermining), and but he also argues for the plausibility of a local PMI, viz., M-PMI, and contends that there are clear advantages of PMI-M over PMI. What I wish to note here is that an additional advantage of PMI-M, or local pessimistic induction generally speaking, is that whereas global PMI dissolves upon a material analysis, a material account of PMI-M does seem viable.

6. Conclusion

I have argued that historical inductions such as the (global) PMI and the problem of unconceived alternatives dissolve if we work with the material theory of induction. The reason is that we lack the material postulates needed to license the pessimistic inference: the great heterogeneity of case studies from the history of science of conceiving, constructing, and discovering (explanatory and predictively successful) theories, along with abundant variety of context that scientists find themselves in and traits that they exhibit, make it unlikely that any commonality will be found strong enough to authorize the induction. One may of course object: so much worse for the material theory of induction! This is a fair point, but there is a more general moral to consider. In various situations one may be able to appeal to the notion of "induction" without much being at stake, but in the context of historical inductions like the PMI and problem of unconceived alternatives "induction" is doing a lot of (philosophically) heavy lifting and so the situation rightful calls for scrutiny. Such scrutiny has led to the various discussed criticism that are presented in the context of more traditional, non-material theories of induction. Accordingly, it seems appropriate to show that—even if we assume randomly sampled historical evidence from the correct period of history and with the proper unit of analysis that is not biased or cherry-picked, with no statistical error, etc.—historical inductions do not fare well on the material side of things. I leave objections to the effect that one ought to construe the PMI as a deductive argument, or through a different framework for induction, e.g., via hypothetical or probabilistic induction, for future work.

References

Chakravartty, A. 2008. "What You Don't Know Can't Hurt You: Realism and the Unconceived." *Philosophical Studies* 137: 149–158.

Crombie, A. C., 1995. *Styles of Scientific Thinking in the European Tradition*, 3 vols. London: Duckworth.

Devitt, M. 2011. "Are Unconceived Alternatives a Problem for Scientific Realism?" *Journal for General Philosophy of Science* 42: 285–293.

Fahrbach, L. 2011. "How the Growth of Science Ends Theory Change." *Synthese* 180: 139–155.

Hacking, I. 1992. "'Style' for historians and philosophers." *Studies in History and Philosophy of Science*, 23(1), 1–20.
Ioannidis, J. P. A. 2005a. "Contradicted and Clinically Stronger Effects in Highly Cited Clinical Research." *Journal of the American Medical Association* 294: 218–228.
Ioannidis, J. P. A. 2005b. "Why Most Published Research Findings Are False."*PLoS Medicine 2*: 696–701.
Kwa, C. 2011. *Styles of Knowing: A New History of Science from Ancient Times to the Present*. Pittsburgh: University of Pittsburgh Press.
Lange, M. 2002. "Baseball, Pessimistic Inductions, and the Turnover Fallacy." *Analysis* 62: 281–285.
Laudan, L. 1981. "A Confutation of Convergent Realism." *Philosophy of Science* 48: 19–49.
Lewis, P. J. 2001. "Why the Pessimistic Induction Is a Fallacy." *Synthese* 129: 371–380.
Magnus, P. D., and C. Callender. 2004. "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71: 320–338.
Mill, J. S. [1872] 1916. *A System of Logic: Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. 8th ed. London: Longman, Green, and Co.
Mizrahi, M. 2013. "The Pessimistic Induction: A Bad Argument Gone too Far." *Synthese* 190:3209–3226.
Mizrahi, M. 2015. "Historical Inductions: New Cherries, Same Old Cherry-picking." *International Studies in the Philosophy of Science* 29: 129–148.
Mizrahi, M. 2016. "The history of Science as a Graveyard of Theories: A Philosophers' Myth?" *International Studies in the Philosophy of Science* 30: 263-278.
Norton, J. D. 2003. "A Material Theory of Induction." Philosophy of Science 70: 647–670.
Norton, J. D. Manuscript. *The Material Theory of Induction*. See http://www.pitt.edu/~jdnorton/papers/material_theory/material.html
Park, S. 2011. "A Confutation of the Pessimistic Induction." *Journal for General Philosophy of Science* 42: 75–84.
Poincaré, H. [1902] 1952. *Science and Hypothesis*. New York: Dover. Originally published as *La science et l'hypothèse*. Paris: Flammarion.
Putnam, H. 1978. Meaning and the Moral Sciences. London: Routledge and Kegan Paul.
Psillos, S.: 1996, 'Scientific Realism and the 'Pessimistic Induction' ', *Philosophy of Science* 63 (Proceedings), S306–S314.
Psillos, S. 1999. Scientific Realism: How Science Tracks Truth. London: Routledge.
Rescher, N. 1987. Scientific Realism: A Critical Reappraisal. Dordrecht: D. Reidel.
Ruhmkorff, S. 2013. "Global and Local Pessimistic Meta-inductions." *International Studies in the Philosophy of Science* 27: 409–428.
Saatsi, J. 2005. "On the Pessimistic Induction and Two Fallacies." *Philosophy of Science* 72: 1088–1098.
Sklar, L. M. (2003). "Dappled theories in a uniform world." *Philosophy of Science*, 70, 424–441.

Stanford, P. K. 2001. “Refusing the Devil’s Bargain: What Kind of Underdetermination Should We take Seriously?” *Philosophy of Science* 68 (Proceedings): S1–S12.

Stanford, P. K. 2006. Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford: Oxford University Press.

Wray, K. Brad. 2015. “Pessimistic Inductions: Four Varieties.” *International Studies in the Philosophy of Science* 29: 61–73.

To be presented at the *2018 PSA Meeting*:

Can Quantum Thermodynamics Save Time?

Noel Swanson*

Abstract

The *thermal time hypothesis (TTH)* is a proposed solution to the problem of time: every statistical state determines a thermal dynamics according to which it is in equilibrium, and this dynamics is identified as the flow of physical time in generally covariant quantum theories. This paper raises a series of objections to the TTH as developed by Connes and Rovelli (1994). Two technical challenges concern the implementation of the TTH in the classical limit and the relationship between thermal time and proper time. Two more conceptual problems focus on interpreting the flow of time in non-equilibrium states and the lack of gauge invariance.

1 Introduction

In both classical and quantum theories defined on fixed background spacetimes, the physical flow of time is represented in much the same way. Time translations correspond to a continuous 1-parameter subgroup of spacetime symmetries, and the dynamics are implemented either as a parametrized flow on statespace (Schödinger picture) or a parametrized group of automorphisms of the algebra of observables (Heisenberg picture). In generally

*Department of Philosophy, University of Delaware, 24 Kent Way, Newark, DE 19716, USA, nswanson@udel.edu

covariant theories, where diffeomorphisms of the underlying spacetime manifold are treated as gauge symmetries, this picture breaks down. There is no longer a canonical time-translation subgroup at the global level, nor is there a gauge-invariant way to represent dynamics locally in terms of the Schrödinger or Heisenberg pictures. Without a preferred flow on the space of states representing time, the standard way to represent physical change via functions on this space taking on different values at different times, also fails. This is the infamous *problem of time*.

Connes and Rovelli (1994) propose a radical solution to the problem: the flow of time (not just its direction) has a thermodynamic origin. Equilibrium states are usually defined with respect to a background time flow (e.g., dynamical stability and passivity constraints reference a group of time translations). Conversely, given an equilibrium state one can derive the dynamics according to which it is in equilibrium. Rovelli (2011) exploits this converse connection, arguing that in a generally covariant theory, *any* statistical state defines a notion of time according to which it is an equilibrium state. The *thermal time hypothesis (TTH)* identifies this state-dependent thermal time with physical time. Drawing upon tools from Tomita-Takesaki modular theory, Connes and Rovelli demonstrate how the TTH can be rigorously implemented in generally covariant quantum theories.

The idea is an intriguing one that, to date, has received little attention from philosophers.[1] This paper represents a modest initial attempt to sally forth into rich philosophical territory. Its goal is to voice a number of technical and conceptual problems faced by the TTH and to highlight some tools that the view has at its disposal to respond.

2 The Thermal Time Hypothesis

We usually think of theories of mechanics as describing the evolution of states and observables through time. Rovelli (2011) advocates replacing this picture with a more general *timeless* one that conceives of mechanics as describing relative correlations between physical quantities divided into two classes, *partial* and *full* observables. Partial observables are quantities that physical measuring devices can be responsive to, but whose value cannot be predicted

[1] Earman (2002), Earman (2011), and Ruetsche (2014) are notable exceptions. Physicists have been more willing to dive in. Paetz (2010) gives an excellent critical discussion of the many technical challenges faced by the TTH.

given the state alone (e.g., proper time along a worldline). A full observable is understood as a coincidence or correlation of partial observables whose value can be predicted given the state (e.g., proper time along a worldline at the point where it intersects another worldline). Only measurements of full observables can be directly compared to the predictions made by the mechanical theory.

A timeless mechanical system is given by a triple $(\mathcal{C}, \Gamma, f)$. $\mathcal{C}$ is the configuration space of partial observables, q^a. A *motion* of the system is given by an unparametrized curve in $\mathcal{C}$, representing a sequence of correlations between partial observables. The space of motions, Γ is the statespace of the system and is typically presymplectic. The evolution equation is given by $f = 0$, where f is a map $f : \Gamma \times \mathcal{C} \to V$, and V is a vector space. For systems that can be modeled using Hamiltonian mechanics, Γ and f are completely determined by a surface Σ in the cotangent bundle $T^*\mathcal{C}$ (the space of partial observables and their conjugate momenta p_a). This surface is defined by the vanishing of some Hamiltonian function $H : T^*\mathcal{C} \to \mathbb{R}$.

If the system has a preferred external time variable, the Hamiltonian can be decomposed as

$$H = p_t + H_0(q^i, p_i, t) \tag{1}$$

where t is the partial observables in $\mathcal{C}$ that corresponds to time. Generally covariant mechanical systems lack such a canonical decomposition. Although these systems are fundamentally timeless, it is possible for a notion of time to emerge thermodynamically. A closed system left to thermalize will eventually settle into a time-independent equilibrium state. Viewed as part of a definition of equilibrium, this thermalization principle requires an antecedent notion of time. The TTH inverts this definition and use the notion of an equilibrium state to select a partial observable in $\mathcal{C}$ as time.

Three hurdles present themselves. The first is providing a coherent mathematical characterization of equilibrium states. The second is finding a method for extracting information about the associated time flow from a specification of the state. Finally, in order to count as an emergent explanation of time, one has to show that the partial observable selected behaves as a traditional time variable in relevant limits.

For generally covariant quantum theories, Connes and Rovelli (1994) propose a concrete strategy to overcome these hurdles. Minimally, such a theory can be thought as a non-commutative C^*-algebra of diffeomorphism-invariant

observables, $\mathfrak{A}$, along with a set of physically possible states, $\{\phi\}$.[2] Via the Gelfand-Nemark-Segal (GNS) construction, each state determines a concrete Hilbert space representation $(\pi_\phi(\mathfrak{A}), \mathcal{H}_\phi)$, and a corresponding von Neumann algebra $\pi_\phi(\mathfrak{A})''$, defined as the double commutant of $\pi_\phi(\mathfrak{A})$.

Connes and Rovelli first appeal to the well-known *Kubo-Martin-Schwinger (KMS) condition* to characterize equilibrium states. A state, ρ, on a von Neumann algebra, $\mathfrak{M}$, satisfies the KMS condition for inverse temperature $0 < \beta < \infty$ with respect to a 1-parameter group of automorphisms, $\{\alpha_t\}$, if for any $A, B \in \mathfrak{M}$ there exists a complex function $F_{A,B}(z)$, analytic on the strip $\{z \in \mathbb{C} | 0 < \mathrm{Im} z < \beta\}$ and continuous on the boundary of the strip, such that

$$\begin{aligned} F_{A,B}(t) &= \rho(\alpha_t(A)B) \\ F_{A,B}(t + i\beta) &= \rho(B\alpha_t(A)) \end{aligned} \tag{2}$$

for all $t \in \mathbb{R}$. The KMS condition generalizes the idea of an equliibirum state to quantum systems with infinitely many degrees of freedom. KMS states are stable, passive, and invariant under the dynamics, $\{\alpha_t\}$. Moreover in the finite limit, the KMS condition reduces to the standard Gibbs postulate.

Although the KMS condition is framed relative to a chosen background dynamics, according to the main theorem of *Tomita-Takesaki modular theory*, every faithful state determines a canonical 1-parameter group of automorphisms according to which it is a KMS state. Connes and Rovelli go on to identify the flow of time with the flow of this state-dependent *modular automorphism group*.

In the GNS representation $(\pi_\phi(\mathfrak{A}), \mathcal{H}_\phi)$, the defining state, ϕ, is represented by a cyclic vector $\Phi \in \mathcal{H}_\phi$. If ϕ is a *faithful* state (i.e., if $\phi(A^*A) = 0$ entails that $A = 0$) then the vector Φ is also separating. In this setting we can apply the tools of Tomita-Takesaki modular theory. The main theorem asserts the existence of two unique modular invariants, an antiunitary operator, J, and a positive operator, Δ. (Here we will only be concerned with the latter.) The 1-parameter family, $\{\Delta^{is} | s \in \mathbb{R}\}$, forms a strongly continuous unitary group,

$$\sigma_s(A) := \Delta^{is} A \Delta^{-is} \tag{3}$$

for all $A \in \pi(\mathfrak{A})''$, $s \in \mathbb{R}$. The defining state is invariant under the flow of the modular automorphism group, $\phi(\sigma_s(A)) = \phi(A)$. Furthermore, $\phi(\sigma_s(A)B) =$

[2]See Brunetti et al. (2003) for a formal development of this basic idea.

$\phi(B\sigma_{s-i}(A))$. Thus ϕ satisfies the KMS condition relative to $\{\sigma_s\}$ for inverse temperature $\beta = 1$.

For any faithful state, this procedure identifies a partial observable, the thermal time, $t_\phi := s$, parametrizing the flow of the (unbounded) thermal hamiltonian $H_\phi := -\ln \Delta$, which has Φ as an eigenvector with eigenvalue zero. We can then go on to decompose the timeless Hamiltonian $H = p_{t_\phi} + H_\phi$. Associated with any such state, there is a natural "flow of time" according to which the system is in equilibrium. But in what sense does this thermal time flow correspond to various notions of physical time? In particular, how is thermal time related to the proper time measured by a localized observer?

Although they do not establish a general theorem linking thermal time to proper time, Connes and Rovelli do make substantial progress on the third hurdle in one intriguing special case. For a uniformly accelerating, immortal observer in Minkowski spacetime, the region causally connected to her worldline is the *Rindler wedge*. In standard coordinates we can explicitly write the observer's trajectory as

$$\begin{aligned} x^0(\tau) &= a^{-1}\sinh(\tau) \\ x^1(\tau) &= a^{-1}\cosh(\tau) \\ x^2(\tau) &= x^3(\tau) = 0 \end{aligned} \tag{4}$$

where τ is the observer's proper time. The wedge region is defined by the condition $x^1 > |x^0|$. The *Bisognano-Wichmann theorem* then tells us that in the vacuum state, the modular automorphism group for the wedge implements wedge-preserving Lorentz boosts — Δ^{is} is given by the boost $U(s) = e^{2\pi i s K_1}$ (where K_1 is the representation of the generator of an x^1-boost). Since the Lorentz boost $\lambda(a\tau)$ implements a proper time translation along the orbit of an observer with acceleration a, $U(\tau) = e^{ai\tau K_1}$ can be viewed as generating evolution in proper time. Comparing these two operators, we find that proper time is directly proportional to thermal time,

$$s = \frac{2\pi}{a}\tau \tag{5}$$

The Unruh temperature measured by the observer is $T = a/2\pi k_b$ (where k_b is Boltzmann's constant), this leads Connes and Rovelli to propose that the Unruh temperature can be interpreted as the ratio between thermal and proper time. Not only does this relationship hold along the orbits of constant

acceleration, but if an observer constructs global time coordinates for the wedge via the process of Einstein synchronization, this global time continues to coincide with the rescaled thermal time flow.

We can now summarize the main content of the TTH:

Thermal Time Hypothesis (Rovelli-Connes)**.** *In a generally covariant quantum theory, the flow of time is defined by the state-dependent modular automorphism group. The Unruh temperature measured by an accelerating observer represents the ratio between this time and her proper time.*

This is a bold idea with a numerous potential implications for quantum physics and cosmology. Over the next three sections, we will consider a series of technical and conceptual objections to the TTH.

3 Thermal Time and Proper Time

The Bisognano Wichmann theorem only applies to immortal, uniformly accelerating observers in the vacuum state of a quantum field theory in flat spacetime. How can we characterize the relationship between thermal and proper time for a broader, more physically realistic class of observers and theories?

A uniformly accelerating mortal observer has causal access to a different region of Minkowski spacetime, the *doublecone* formed by the intersection of her future lightcone at birth and her past lightcone at death. Because wedges and doublecones can be related by a conformal transformation, in conformally invariant theories, geometric results from wedge algebras can be transferred onto the doubelcone algebras. In the vacuum state of a conformal theory, the doublecone modular automorphism group acts as Hislop-Longo transformations (Hislop and Longo, 1982). Martinetti and Rovelli (2003) use this result to calculate the corresponding relationship between thermal time and proper time for a uniformly accelerating mortal observer:

$$s = \frac{2\pi}{La^2}(\sqrt{1 + a^2L^2} - \cosh a\tau) \tag{6}$$

where L is the observer's lifetime. (The relationship is more complicated in this case due to the fact that proper time is bounded while modular time is unbounded.) For most of the observer's lifespan, s is an approximately constant function of τ, allowing the Unruh temperature to again be interpreted as the local ratio between thermal and proper time.

This is the best we can hope for. Trebels (1997) proves that arbitrary doublecone automorphisms act as local dynamics, only if they act as scaled Hislop-Longo transformations.[3] Of course, if nature is described by a non-conformal theory, then there is no guarantee that the doublecone modular automorphisms will have a suitable geometric interpretation. Saffary (2005) goes further, arguing that they will not have geometric significance in any theory with massive particles. The mathematical results backing this conjecture, however, are only partial.[4]

Attempting to generalize the TTH to cover non-uniform acceleration and non-vacuum states generates further difficulties. Work on the Unruh effect for non-uniformly accelerating observers (e.g., Jian-yang et al. 1995), indicates that such observers feel an acceleration-dependent thermal bath, reflecting the shifting ratio between constant thermal time and acceleration-dependent proper time. The TTH must explain the phenomenological experience of the observer who will presumably age according to her proper time, not the background thermal time flow. On top of this, if the global state is not a vacuum state, then it is not clear that the wedge modular automorphisms will carry a dynamical interpretation at all. The Radon-Nikodym theorem ensures that the action of the modular automorphism group uniquely determines the generating state. If ϕ, ψ are two (faithful, normal) states on a von Neumann algebra $\mathfrak{M}$, then the associated modular automorphism groups σ^t_ϕ, σ^t_ψ differ by a non-trivial inner automorphism, $\sigma^t_\phi(A) = U\sigma^t_\psi(A)U^*$, for all $A \in \mathfrak{M}$, $t \in \mathbb{R}$, so the general wedge dynamics will not be simple rescalings of the vacuum case.

None of these are knockdown objections since so little is known about the geometric action of modular operators apart from the Bisognano-Wichmann theorem and its conformal generalization. But our current ignorance also presents a major challenge. (The situation is even less clear in general curved spacetime settings.) The defender of the TTH has at least four options on

[3]Formally, Trebels requires that local dynamics be continuous 1-parameter groups of automorphisms of the doublecone algebra that preserve subalgebra localization as well as spacelike and timelike relations between interior points. For a detailed discussion of Trebels's results, see Borchers (2000), §3.4.

[4]In the massless case, the modular generators are ordinary differential operators, δ_0, of order 1. In the massive case, it has been conjectured that the modular generators are pseudo-differential operators $\delta_m = \delta_0 + \delta_r$, where the leading term is given by the massless generator δ_0 and δ_r is a pseudo-differential operators of order < 1. This second term is thought to give rise to non-local action without geometric interpretation.

the table.

She can hold out hope for a suitably general dynamical interpretation of modular automorphisms in a wide class of physically significant states. There is some indication that states of compact energy (e.g., states satisfying the Döplicher-Haag-Roberts and Buchholz-Fredenhagen selection criteria) give rise to well-behaved modular structure on wedges. In this case the wedge modular automorphisms can be related to those in the vacuum state by the Radon-Nikodym derivative (Borchers, 2000). The analogous problem for doublecones is still open.

Alternatively, she could reject the idea that the thermal time flow determines the temporal metric directly. Thermal time would only give rise to the order, topological, and group theoretic properties of physical time. Metrical properties would be determined by a completely different set of physical relations. Some support for this idea comes from the justification of the clock hypothesis in general relativity. Rather than stipulating the relationship between proper time, τ, and the length of a timelike curve $||\gamma||$, Fletcher (2013) shows that for any $\epsilon > 0$, there is an idealized lightclock moving along the curve which will measure $||\gamma||$ within ϵ. This justifies the clock hypothesis by linking the metrical properties of spacetime to the readings of tiny lightclocks. If the metrical properties of time experienced by localized observers arises via some physical mechanism akin to light clock synchronization. This would explain why the duration of time felt by the observer matches her proper time and not the geometrical flow of thermal time.

Perhaps motivated by the justification of the clock hypothesis, the defender of the TTH could attempt to argue that the metrical properties of time emerge from modular dynamics in the short distance limit of the theory. If the theory has a well-defined ultraviolet limit, the renormalization group flow should approach a conformal fixed point. Buchholz and Verch (1995) prove that in this limit, the double-cone modular operators act geometrically like wedge operators implementing proper time translations along the observer's worldline. It is unlikely that the physics at this scale would directly impact phenomenology, but the asymptotic connection might turn out to be important for explaining the metrical properties of spacetime (which bigger, more realistic lightclocks measure) as emergent features of some underlying theory of quantum gravity.

A final option would be to go back to the drawing board. Rovelli and Connes briefly note that since the modular automorphisms associated with each (faithful, normal) state of a von Neumann algebra are connected by

inner automorphisms, they all project down onto the same 1-parameter group of outer automorphisms the algebra. The TTH could be revised to claim that this canonical state-independent flow represents the non-metrical flow of physical time. It is not known, however, under what circumstances the outer flow acts in suitably geometric fashion to be interpretable as local dynamics, so it remains to be seen whether or not this is a viable option. The move does have immediate consequences for the global dynamics, however. Since the global algebra is expected to be type I, all modular automorphisms will be inner. As a result the canonical group of outer automorphisms is trivial. At a global level, there is no passage of time. At the local level, time emerges as a consequence of our ignorance of the global state.

4 The Classical Limit

The classical limit presents a different kind of challenge. Conceptually, nothing about the idea that a statistical state selects a preferred thermal time requires that the theory be quantum mechanical. The proposed mechanism for selecting a partial observable using modular theory, however, does appear to rely on the noncommutativity of quantum observables. If we model classical systems using abelian von Neumann algebras, then every state is tracial (i.e., $\phi(AB) = \phi(BA)$), and consequently every associated modular automorphism group acts as the identity, trivializing the thermal time flow. Does the TTH have a classical counterpart, or is quantum mechanics required to save time in a generally covariant setting?

Arguing by analogy with standard quantization procedures, Connes and Rovelli suggest that in the classical limit commutators need to be replaced by Poisson brackets. We begin with an arbitrary statistical state, ρ, represented by a probability distribution over a classical statespace Γ:

$$\int_{\Gamma} dx\ \rho(x) = 1 \tag{7}$$

where $x \in \Gamma$ is a timeless microstate. By analogy with the Gibbs postulate, we can introduce the "thermal Hamiltonian,"

$$H_\rho = -\ln \rho \tag{8}$$

With respect to the corresponding Hamiltonian vector field, the evolution of

an arbitrary classical observable, $f \in C^\infty(\Gamma)$, is given by

$$\frac{d}{ds} f = \{-\ln \rho, f\} \tag{9}$$

and $\rho = \exp(-H_\rho)$. With respect to the Poisson bracket structure, the classical algebra of observables is non-abelian. Gallavotti and Pulvirenti (1976) use this non-abelian structure to define an analogue of the KMS condition. Is this connection strong enough to support a version of the TTH in ordinary general relativity? Or does it only serve to aid us in understanding how the thermal time variable behaves in the transition from quantum theory to classical physics?

The difficulty lies in connecting the thermal time flow for an arbitrary statistical state to our ordinary conception of time. In the quantum case this link was provided by the Bisognano-Wichmann theorem, which does not have a classical analogue. The problem is magnified by the lack of a full understanding of statistical mechanics and thermodynamics in curved spacetime. Rovelli has done some preliminary work on developing a full theory of generally covariant thermodynamics based on the foundation supplied by the TTH, including an elegant derivation of the Tolman-Ehrenfest effect, but the field is still young.[5]

Setting aside these broader interpretive challenges for now, an important first step lies in obtaining a better understanding the classical selection procedure outlined above. As it turns out, the commutator-to-Poisson-bracket ansatz is on firmer foundational footing than one might initially suspect. As emphasized by Alfsen and Shultz (1998), non-abelian C^*-algebras have a natural *Lie-Jordan structure*:

$$AB = A \bullet B - i(A \star B) \ , \tag{10}$$

The non-associative Jordan product, $\bullet$, encodes information about the spectra of observables, while the associative Lie product, $\star$, encodes the generating relation between observables and symmetries. The significance of the commutator, is that it defines the canonical Lie product, $A \star B := i/2[A, B]$. Classical mechanical theories formulated on either a symplectic or Poisson manifold have a natural Lie-Jordan structure as well. The standard product of functions defines an associative Jordan product, encoding spectral information, while the Poisson bracket determines the associative Lie product,

[5] See Rovelli and Smerlak (2011).

describing how classical observables generate Hamiltonian vector fields on statespace. Together, this structure is called a *Poisson algebra*. The primary difference between the classical and quantum cases is the associativity/non-associativity of the Jordan product.

These considerations point towards the idea that the appropriate classical analogue of a noncommutative von Neumann algebra, is not a commutative von Neumann algebra, but a Poisson algebra. In this setting, initial strides towards a classical analogue of modular theory have been made by Weinstein (1997). Given any smooth density, μ, on a Possion manifold, Γ, Weinstein defines a corresponding *modular vector field* ϕ_μ given by the operator $\phi_\mu : f \to \mathrm{div}_\mu H_f$ where H_f is the Hamiltonian vector field associated with a classical observable, $f \in C^\infty(\Gamma)$. The antisymmetry of the Poisson bracket entails that the operator ϕ_μ is a vector field on Γ. Weinstein proposes ϕ_μ as the classical analogue of the modular automorphism group. It characterizes the extent to which the Hamiltonian vector fields are divergence free (with respect to the density μ), vanishing iff all Hamiltonian vector fields are divergence free.

We can connect Weinstein's classical modular theory to the TTH. If Γ is a symplectic manifold and we let μ be the density associated with the canonical Liouville volume form, then $\phi_\mu(f) = 0$ for all observables. This reflects the conservation of energy by Hamiltonian flows in symplectic dynamical systems. Given any statistical state, however, we can define an associated density which leads to a nontrivial modular vector field. For any positive function, h, we have

$$\phi_{h\mu} = \phi_\mu + H_{-\ln h} = H_{-\ln h}. \tag{11}$$

Therefore any statistical state, ρ, defines a modular vector field equivalent to the Hamiltonian vector field $H_{-\ln \rho}$ associated with the density $e^{-\ln \rho}\mu$. We immediately recognize $-\ln \rho$ as the thermal Hamiltonian postulated by Connes and Rovelli. Clearly, $e^{is \ln \rho} \rho e^{-is \ln \rho} = \rho$, thus the state is invariant with respect to the flow of $H_{-\ln \rho}$. Additionally, it can be shown that ρ satisfies the KMS condition with respect to these dynamics, hence, from the perspective of the associated time flow ρ resembles an invariant equilibrium state just as in the quantum case.

5 Conceptual Challenges

As we have seen in the previous two sections, the TTH faces a number of technical challenges (some of which look easier to overcome than others). There are, however, several deeper conceptual problems looming in the background which pose a more serious challenge to the viability of the hypothesis. Here, we will discuss two of the most pressing.

The first, which we will call the *generality problem*, draws upon the preceding discussion of the classical limit. While mathematically speaking, Weinstein's modular vector field gives us a method for selecting a canonical thermal time flow in a classical theory, physical speaking, there is no reason why we should view the corresponding thermal time as physical time. As we have seen, any statistical state determines thermal dynamics according to which it is a KMS state, however, if ρ is a non-equilibrium state, the resultant thermal time flow does not align with our ordinary conception of time. By the lights of thermal time, a cube of ice in a cup of hot coffee is an invariant equilibrium state! The same problem arises in the quantum domain — only for states which are true equilibrium states will the thermal time correspond to physical time.

It appears inevitable that the TTH will have to be tempered. Rather than letting any state determine a corresponding flow of thermal time, only certain reference states should be permitted. Apart from the problem of providing an intrinsic, non-dynamical characterization of such states, if a system is not in one of these, it is hard to envision how a counterfactual state of affairs can determine the actual flow of time.[6] This might provide more reasons for the defender of the TTH to explore the state independent, outer modular flow. Alternatively, she could try to argue that local non-equilibrium behavior can be viewed as small fluctuations from some background state. On this approach, the local flow of time in my office according to which the ice

[6] A closely related worry, what we might call the *background-dependence problem*, has been voiced by Earman (2011) and Ruetsche (2014). Their concern is that we can only identify modular automorphisms as dynamics because we already have a rich spatiotemporal geometry in the background. This casts doubt on whether the TTH can provide a coherent definition of time in situations where such structure is absent (as required to solve the full problem of time). This is exacerbated if the TTH is modified in response to the generality problem. Unless the modular automorphism group can always be viewed dynamically, the defender of the TTH will be hard pressed to find constraints capable of separating the dynamical cases from the non-dynamical cases which are independent of all background temporal structure.

melts and the coffee cools is not defined by the thermal state of the ice/coffee system, but the thermal state of some larger enveloping system (the entire universe perhaps). Rovelli (1993) hints in this direction, calculating that in a Friedman-Robertson-Walker universe, the thermal time induced by the equilibrium state of the cosmic microwave background will be proportional to the FRW time. While the connection is intriguing, it seems unlikely that an explanation of this sort will be able to account for the flow of time experienced by localized, mortal observers like us. It would be truly remarkable to discover that our faculties of perception are sensitive to the thermal features of the CMB.

The second problem is the *gauge problem*. The TTH does succeed in providing a means to select a privileged 1-parameter flow on the space of full, gauge invariant observables of a generally covariant theory. What makes this flow interpretable as a *dynamical* flow, however, is its description as a sequence of correlations between partial observables. The difficulty is that these partial observables are not diffeomorphism invariant. Assuming that we treat diffeomorphisms in generally covariant theories as standard gauge symmetries (which is how we got into the problem of time in the first place), then the partial observables are just descriptive fluff. They do not directly represent physical features of our world.

The problem is *not* the resultant timelessness of fundamental physics. The TTH adopts this dramatic conclusion willingly. The problem is that the TTH is supposed to explain how the appearance of time and change emerge from timeless foundations. But the explanation given is couched in gauge-dependent language, and it is not apparent how we can extract a gauge invariant story from it. We can introduce partial observables and use correlations between them to calculate and predict emergent dynamical behavior, but we cannot use these correlations to *explain* that behavior. We lack a gauge invariant picture of generally covariant theories, and the TTH, at least in its present form, does not provide one.

Can a revised TTH give us the explanatory tools needed to understand the flow of time without reference to partial observables, or, does the entire framework of timeless mechanics require us to revise our conception of how ontology, explanation, and gauge symmetries are related?[7] Whether or not

[7]Drifting in the latter direction, Rovelli (2014) suggests that gauge-dependent quantities are more than just mathematical redundancies, "they describe handles through which systems couple: they represent real relational structures to which the experimentalist has access in measurement by supplying one of the relata in the measurement procedure itself."

quantum thermodynamics can save time may rest on the solutions to these new incarnations of vexingly familiar philosophical problems.

References

Alfsen, E. and F. Shultz (1998). Orientation in operator algebras. *Proceedings of the National Academy of Sciences, USA 95*, 6596–6601.

Borchers, H. J. (2000). On revolutionizing quantum field theory with Tomita's modular theory. *Journal of Mathematical Physics 41*(6), 3604–3673.

Brunetti, R., K. Fredenhagen, and R. Verch (2003). The generally covariant locality principle – a new paradigm for local quantum field theory. *Communications in Mathematical Physics 237*, 31–68.

Buchholz, D. and R. Verch (1995). Scaling algebras and renormalization group in algebraic quantum field theory. *Reviews in Mathematical Physics 7*, 1195.

Connes, A. and C. Rovelli (1994). Von Neumann algebra automorphisms and time-thermodynamics relation in generally covariant quantum theories. *Classical and Quantum Gravity 11*(12), 2899.

Earman, J. (2002). Thoroughly modern McTaggart. *Philosopher's Imprint*, 2. http://www.philosophersimprint.org/002003/.

Earman, J. (2011). The Unruh effect for philosophers. *Studies in History and Philosophy of Modern Physics 42*, 81–97.

Fletcher, S. (2013). Light clocks and the clock hypothesis. *Foundations of Physics 43*, 1369–1383.

Gallavotti, G. and M. Pulvirenti (1976). Classical KMS condition and Tomita-Takesaki theory. *Communications in Mathematical Physics 46*, 1–9.

Hislop, P. D. and R. Longo (1982). Modular structure of the local algebras associated with a free massless scalar field theory. *Communications in Mathematical Physics 84*, 71.

Jian-yang, Z., B. Aidong, and Z. Zheng (1995). Rindler effect for a nonuniformly accelerating observer. *International Journal of Theoretical Physics 34*, 2049–2059.

Martinetti, P. and C. Rovelli (2003). Diamond's temperature: Unruh effect for bounded trajectories and thermal time hypothesis. *Classical and Quantum Gravity 20*(22), 4919.

Paetz, T.-T. (2010). An analysis of the 'thermal-time concept' of Connes and Rovelli. Master's thesis, Georg-August-Universität Göttingen.

Rovelli, C. (1993). The statistical state of the universe. *Class.Quant.Grav. 10*, 1567.

Rovelli, C. (2011). Forget time: Essay written for the FQXi contest on the nature of time. *Foundations of Physics*.

Rovelli, C. (2014). Why gauge? *Foundations of Physics 44*(1), 91–104.

Rovelli, C. and M. Smerlak (2011). Thermal time and Tolman–Ehrenfest effect: 'temperature as the speed of time'. *Classical and Quantum Gravity 28*(7), 075007.

Ruetsche, L. (2014). Warming up to thermal the thermal time hypothesis. Quantum Time Conference, University of Pittsburgh, March 28-29.

Saffary, T. (2005). *Modular Action on the Massive Algebra*. Ph. D. thesis, Hamburg.

Trebels, S. (1997). *Über die Geometrische Wirkung Modularer Automoprhismen*. Ph. D. thesis, Göttingen.

Weinstein, A. (1997). The modular automorphism group of a Poisson manifold. *Journal of Geometry and Physics 23*, 379–394.

Neural redundancy and its relation to neural reuse

Abstract

Evidence of the pervasiveness of neural reuse in the human brain has forced a revision of the standard conception of modularity in the cognitive sciences. One persistent line of argument against such revision, however, draws from a large body of experimental literature attesting to the existence of cognitive dissociations. While numerous rejoinders to this argument have been offered over the years, few have grappled seriously with the phenomenon. This paper offers a fresh perspective. It takes the dissociations seriously, on the one hand, while affirming that traditional modularities of mind do not do justice to the evidence of neural reuse, on the other. The key to the puzzle is neural redundancy. The paper offers both a philosophical analysis of the relation between reuse and redundancy, as well as a plausible solution to the problem of dissociations.

1. Introduction

Cognitive science, linguistics and the philosophy of psychology have long been under the spell of "the modularity of mind" (Fodor 1983), or the idea of the mind as a modular system (see e.g. de Almeida and Gleitman 2018). In contemporary psychology, a modular system is generally understood to be "one consisting of functionally specialized subsystems responsible for processing different classes of input (e.g. for vision, hearing, human faces, etc.), or at any rate for handling specific cognitive tasks" (Zerilli 2017a, 231). According to this theory, "human cognition can be decomposed into a number of functionally independent processes, [where] each of these processes operates over a distinct domain of cognitive information" (Bergeron 2007, 176). What makes one process distinguishable from another is its "functional independence, the fact that one can be affected, in part or in totality, without the other being affected, and vice versa" (Bergeron 2007, 176). Furthermore, given that functional processes are realized in the brain, a functionally specialized process is one which presumably occupies a distinctive portion of neural tissue, though not necessarily a small, closely circumscribed and contiguous region. So fruitful and influential has this model been that it is safe to say that in many quarters of the cognitive sciences—and most especially in cognitive psychology, cognitive neuropsychology and evolutionary psychology—modularity is essentially the received view (McGeer 2007; Carruthers 2006; de Almeida and Gleitman 2018).

Seattle, WA; 1-4 November 2018 -527-

Developments in cognitive neuroscience over the past thirty years, however, have discomfited the modular account. More evidence than ever before points to the pervasiveness of neural reuse in the human brain—the "redeployment" or "recycling" of neural circuits over widely disparate cognitive domains (Anderson, 2010, 2014; Dehaene, 2005). As the terminology suggests, theories of "re-use" posit the "exaptation" of established and diachronically stable neural circuits over the course of evolution or normal development *without* loss of original function, so that the functional contribution of a circuit is preserved across multiple task domains.[1] As Anderson (2010, 246) explains, "rather than posit a functional architecture for the brain whereby individual regions are dedicated to large-scale cognitive domains like vision, audition, language and the like, neural reuse theories suggest that low-level neural circuits are used and reused for various purposes in different cognitive and task domains." According to the theory, just the same circuits exapted for one purpose can be exapted for another provided sufficient intercircuit pathways exist to allow alternative arrangements of them. Indeed, the same parts put together in *different* ways will yield different functional outcomes, just as "if one puts together the same parts *in the same way* one will get the same functional outcomes" (Anderson 2010, 247, my emphasis). The evidence here converges from heterogeneous sources and research paradigms, including neuroimaging (Anderson 2007a; 2007b; 2007c; 2008), computational (Eliasmith 2015), biobehavioral (Casasanto and Dijkstra 2010) and interference paradigms (Gauthier et al.

[1] This usage of "exaptation" is somewhat misleading, since exaptation usually implies loss of original function (see Godfrey-Smith 2001).

2003), and exempts practically no area of the brain (Leo et al. 2012, 2), including areas long regarded as specialized hubs for certain types of sensory processing, e.g. visual and auditory pathways (Striem-Amit and Amedi 2014). Among other things, this means that one of the hallmark features of a module—its domain specificity (Coltheart 1999)—looks too stringent a requirement to prove useful.[2] For neural reuse demonstrates that any one module will typically be sensitive to *more* than one stimulus, including—most importantly—those channeled along intermodal pathways. Meanwhile efforts to salvage a computational or "software" theory of modularity, which carries no commitments regarding implementation, have met with scepticism (Anderson 2007c; 2010; Anderson & Finlay 2014) if not outright opposition (Zerilli 2017a).[3] And while the brain could still be modular in some other sense, what is clear is that the strict domain-specific variety of modularity can no longer serve as an appropriate benchmark.[4]

And yet there is a persistent line of argument *against* this conclusion which draws from a large body of experimental literature attesting to the existence of cognitive

[2] The sense of domain specificity that is relevant here refers to a module's sensitivity to a restricted class of inputs as defined by a domain of psychology—such as visual, auditory or linguistic information. For discussion of alternative senses, see Barrett and Kurzban (2006) and Prinz (2006).

[3] Though by no means universally (see e.g. Carruthers 2010; Jungé and Dennett 2010).

[4] Nor, for that matter, can its cognate property, informational encapsulation (see below).

Seattle, WA; 1-4 November 2018 -529-

dissociations, in which a cognitive ability (say language) is either selectively impaired (linguistic ability is compromised, but no other cognitive ability seems to be materially affected) or selectively spared (general intelligence is compromised, while linguistic abilities function more or less as they should). This literature, most vividly exemplified in lesion studies, is frequently cited in support of classical modularities of mind—be they inspired by the likes of Jerry Fodor (1983), evolutionary psychology (e.g. Cosmides and Tooby 1994; Barrett and Kurzban 2006; Carruthers 2006) or some variation thereof (e.g. ACT-R). While numerous rejoinders to this line of thinking have been offered over the years, few have grappled seriously with the phenomenon, either dismissing the dissociations as noisy, or reasoning from architectural considerations that even nonmodular systems can generate dissociations (Plaut 1995). The aim of this paper is to offer a fresh perspective on this vexed topic. I take the dissociation evidence seriously, on the one hand, while affirming that traditional modularities of mind do not do justice to the evidence of neural reuse, on the other. I do this by invoking neural redundancy, an important feature of cortical design that ensures we have various copies of the same elementary processing units that can be put to alternative (if computationally related) uses in enabling diverse cognitive functions. In the course of the discussion I offer a philosophical explication of the relationship between neural reuse and neural redundancy.

2. What is the Problem? Cognitive Dissociations and Neural Reuse

Let us take an especially contentious question to underscore the nature of the problem we are dealing with and how redundancy might assist in its illumination. The question is this: Does language rely on specialized cognitive and neural machinery, or does it rely on the same machinery that allows us to get by in other domains of human endeavour? The question is bound up with many other questions of no less importance, questions concerning the uniqueness of the human mind, the course of biological evolution and the power of human culture. What is perhaps a little unusual about this question, however—unusual for a question whose answer concerns both those working in the sciences and the humanities—is that it can be phrased as a polar interrogative, i.e. as a question which admits of a yes or no response. And indeed the question has divided psychologists, linguists and the cognitive science community generally for many decades now, more or less into two camps. I would like to sketch the beginnings of an answer to this question—and others like it—in a way that does not pretend it can receive a simple yes or no response.

First of all, let me stress again that neural reuse is as well verified a phenomenon as one can expect in the cognitive sciences, and that it has left virtually no domain of psychology untouched. Neural reuse suggests that there is nothing so specialized in the cortex that it cannot be repurposed to meet new challenges while retaining its capacity for meeting old ones. In that regard, to be sure, what I am proposing is unapologetically on the side of those who maintain that language, as well as many other psychological capacities, are

not cognitively special—e.g. that there is no domain-specific "language organ" (cf. Chomsky 1980,39, 44; 1988, 159; 2002, 84-86).

And yet I would like to carefully distinguish this claim from the claim that there are no areas of the brain that subserve exclusively linguistic functions. The neuropsychological literature offers striking examples of what appear to be fairly clean dissociations between linguistic and nonlinguistic capacities, i.e. cases in which language processing capacities appear to be disrupted without impeding other cognitive abilities, and cases in which the reverse situation holds (Fedorenko et al. 2011; Hickok and Poeppel 2000; Poeppel 2001; Varley et al. 2005; Luria et al. 1965; Peretz and Coltheart 2003; Apperly et al. 2006). An example would be where the ability to hear words is disrupted, but the ability to recognize non-word sounds is spared (Hickok and Poeppel 2000; Poeppel 2001). Discussing such cases, Pinker and Jackendoff (2005, 207) add that "[c]ases of amusia and auditory agnosia, in which patients can understand speech yet fail to appreciate music or recognize environmental sounds…show that speech and non-speech perception in fact doubly dissociate." Although dissociations are to some extent compatible with reuse—indeed there is work suggesting that focal lesions can produce specific cognitive impairments within a range of nonclassical architectures (Plaut 1995)—and it is equally true that often the dissociations reported are noisy (Cowie 2008), still their very ubiquity needs to be taken seriously and accounted for in a more systematic fashion than many defenders of reuse have been willing to do (see e.g. Anderson 2010, 248; 2014, 46-48). After all, a good deal of support for

theories of reuse comes from the neuroimaging literature, which is somewhat ambiguous taken by itself. As Fedorenko et al. (2011, 16428) explain:

> standard functional MRI group analysis methods can be deceptive: two different mental functions that activate neighbouring but non-overlapping cortical regions in every subject individually can produce overlapping activations in a group analysis, because the precise locations of these regions vary across subjects, smearing the group activations. Definitively addressing the question of neural overlap between linguistic and nonlinguistic functions requires examining overlap within individual subjects, a data analysis strategy that has almost never been applied in neuroimaging investigations of high-level linguistic processing.

When Fedorenko and her colleagues applied this strategy themselves, they found that "most of the key cortical regions engaged in high-level linguistic processing are not engaged by mental arithmetic, general working memory, cognitive control or musical processing," and they think that this indicates "a high degree of functional specificity in the brain regions that support language" (2011, 16431). While I do not believe that claims of this strength have the least warrant—as I shall explain, functional specificity cannot be established merely by demonstrating that a region is selectively engaged by a task—these results do at least substantiate the dissociation literature in an interesting way and make it more difficult for

Seattle, WA; 1-4 November 2018 -533-

those who would prefer to dismiss the dissociations with a ready-made list of alternative explanations. Similar results were found by Fedorenko et al. (2012).

3. How Might Redundancy Feature In a Solution?

With rare exceptions (e.g. Friston and Price 2003; Barrett and Kurzban 2006; Jungé and Dennett 2010), redundancy has passed almost unnoticed in the philosophical and cognitive science literature. This is in stark contrast to the epigenetics literature, where redundancy and the related concept of degeneracy[5] have been explored to some depth (e.g. see Edelman and Gally 2001; Mason 2010; Whiteacre 2010; Deacon 2010; Iriki and Taoka 2012; Maleszka et al. 2013). The idea behind neural redundancy is that, for good evolutionary reasons (see below), the brain incorporates a large measure of redundancy of function. Brain regions (such as cortical columns and similar structures) fall in an iterative, repetitive and almost lattice-like arrangement in the cortex. Neighbouring columns have similar response properties: laminar and columnar changes are for the most part smooth—not abrupt—as one moves across the cortex, and adjacent modules do not differ markedly from one another in their basic structure and computations (if they really differ at all when taken in such

[5] Redundancy occurs when items have the same structure and function (i.e. are both isomorphic and isofunctional). Degeneracy occurs when items having *different* structures can perform the same function (i.e. are heteromorphic but isofunctional). Degeneracy implies genuine multiple realization (see Zerilli 2017b).

proximity). Regional *solitarity* is therefore not likely to be a characteristic of the brain (Anderson 2014, 141).[6] That is to say, we do not possess just one module for X, and one module for Y, but in effect several *copies* of the module for X, and several copies of the module for Y, all densely stuffed into the same cortical zones. As Buxhoeveden and Casanova (2002, 943) explain of neurons generally:

> In the cortex, more cells do the job that fewer do in other regions....As brain evolution paralleled the increase in cell number, a reduction occurred in the sovereignty of individual neurones; fewer of them occupy critical positions. As a consequence, plasticity and redundancy have increased. In nervous systems containing only a few hundred thousand neurones, each cell plays a more essential role in the function of the organism than systems containing billions of neurones.

The same principle very likely holds for functionally distinct groupings of neurons (i.e. cortical columns and like structures), as Jungé and Dennett (2010, 278) conjecture:

> It is possible that specialized brain areas contain a large amount of structural/computational redundancy (i.e., many neurons or collections of neurons

[6] The term "solitarity" is Anderson's, but while he concedes that solitarity will be "relatively rare," he does not appear to believe that anything particularly significant follows from this. See also Anderson (2010, 296).

> that can potentially perform the same class of functions). Rather than a single neuron or small neural tract playing roles in many high-level processes, it is possible that distinct subsets of neurons within a specialized area have similar competencies, and hence are redundant, but as a result are available to be assigned individually to specific uses....In a coarse enough grain, this neural model would look exactly like multi-use (or reuse).

This is plausibly why capacities which are functionally very closely related, but which for whatever reason are forced to recruit different neural circuits, will often be localized in broadly the same regions of the brain. For instance, first and second languages acquired early in ontogeny settle down in nearly the same region of Broca's area; and even when the second language is acquired in adulthood the second language is represented nearby within Broca's area (while artificial languages are not) (Kandel & Hudspeth 2013). The neural coactivation graphs of such composite networks must look very similar. Indeed these results suggest—and a redundancy model would predict—that two very similar tasks which are forced to recruit different neural circuits should exhibit similar patterns of activation. And this is more or less what we find (see below).

One might be tempted to think that redundancy and reuse pull in opposite directions. This is because whereas reuse posits that neural circuits get reused across different tasks and task categories, redundancy accommodates the likelihood of diverse

cognitive functions being activated by structurally and computationally equivalent circuits running in parallel: instead of a single circuit being reused across domains, two, three or more *copies* of that same circuit may be recruited differentially across those domains, such that no *single* circuit gets literally "re-used." But there is no substantive tension here. The redundancy account in truth *supplements* the reuse picture in a way that is consistent with the neuroimaging data, faithful to the core principle of reuse, and compatible with the apparent modularization and separate modifiability of technical and acquired skills in ontogeny. Evidence of the reuse of neural circuits to accomplish different tasks has, in fact, been adduced in aid of a theory which posits the reuse of the same neural *tokens* to accomplish these different tasks. Redundancy means we must accept that at least some of the time what we may actually be witnessing is reuse of the same *types* to accomplish these tasks. This does not diminish the standing of reuse. Let me explain.[7]

To the extent that a particular composite reuses types, and is dissociable pro tanto—residing in segregated brain tissue that is not active outside the domain in question—it is true that to that extent its constituents will *appear* to be domain-specific. But in this case looks will be deceiving. The classical understanding of domain specificity in effect *assumes* solitarity—that a module for X does something which no other module can do as well, or

[7] For a developmental twist on the type/token distinction invoked in the context of modular theorizing about the mind, see Barrett (2006).

Seattle, WA; 1-4 November 2018 -537-

that even *if* another module can do X as well, taken together these X-ing modules do not perform outside the X-domain. Here is an example of the latter idea (Bergeron 2007, 176):

> a pocket calculator could have four different division modules, one for dividing numbers smaller than or equal to 99 by numbers smaller than or equal to 99, a second one for dividing numbers smaller than or equal to 99 by numbers greater than 99, a third one for dividing numbers greater than 99 by numbers greater than 99, and a fourth one for dividing numbers greater than 99 by numbers smaller than or equal to 99. In such a calculator, these four capacities could all depend on (four versions of) the same algorithm. Yet, random damage to one or more of these modules in a number of such calculators could lead to observable (double) dissociations between any two of these functions.

Here, each module performs fundamentally the same algorithm, but in distinct hardware, such that dissociations are observable between any two functions. Notice, however, that none of these modules performs outside the "division" domain. This is what allows such duplicate modules to be considered domain-specific—they perform functions which, for all that they might run in parallel on duplicate hardware, are unique to a specific domain of operation, in this case division. If such modules could do work outside the division domain, they would lose the status of domain specificity, and acquire the status of domain neutrality (i.e. they would be domain-general). This is why a module that appears dedicated to a

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -538-

particular function may not be domain-specific in the classical sense. Dedication is not the same as domain specificity, and redundancy, whether of calculator algorithms or neural circuits, explains why. A composite of neural regions will be dedicated without being domain-specific if its functional resources are accessible to other domains through the deployment (reuse) of neural surrogates (i.e. redundant or "proxy" tokens). In this case its constituents will be multi-potential but single-use (Jungé & Dennett 2010, 278), and the domain specificity on display somewhat cosmetic. To take an example with more immediate relevance to the brain, a set of cortical columns that are structurally and computationally similar may be equally suited for face recognition tasks, abstract-object recognition tasks, the recognition of moving objects, and so on. One of these columns could be reserved for faces, another for abstract objects, another for moving objects, and so on. What is noteworthy is that while the functional activation may be indistinguishable in each case, and the same *type* of resource will be employed on each occasion, a different *token* module will be at work at any one time. To quote Jungé and Dennett (2010, 278) again:

> In an adult brain, a given neuron [or set of neurons] would be aligned with only a single high-level function, whereas each area of neurons would be aligned with very many different functions.

Such modules (and composites) are for all intents and purposes *qualitatively* identical, though clearly not *numerically* identical, meaning that while they share their properties, they

Seattle, WA; 1-4 November 2018 -539-

are not *one and the same* (Parfit 1984). The evidence of reuse is virtually all one way when it comes to the pervasiveness of functional inheritance across cognitive domains. It may be that this inheritance owes to reuse of the same tokens (literal reuse) or to reuse of the same types (reuse by proxy), but the inheritance itself has been amply attested. This broader notion of reuse still offers a crucial insight into the operations of cognition, and I dare say represents a large part of the appeal of the original massive redeployment hypothesis (Anderson 2007c).

It is interesting to note in this respect that although detractors have frequently pointed out the ambiguity of neuroimaging evidence on account of its allegedly coarse spatial resolution (e.g. Carruthers 2010), suggesting that the same area will be active across separate tasks and task categories even if distinct but spatially adjacent and/or interdigitated circuits are involved in each case, this complaint can have no bearing on reuse by proxy. Fedorenko et al. (2011, 16431) take their neuroimaging evidence to support "a high degree of functional specificity in the brain regions that support language," but their results do not license this extreme claim. The regions they found to have been selectively engaged by linguistic tasks were all adjacent to the regions engaged in nonlinguistic tasks. Elementary considerations suggest that they have discovered a case of reuse by proxy involving language: the domains tested (mental arithmetic, general working memory, cognitive control and musical processing) make use of many of the same computations as high-level linguistic processing, even though they run them on duplicate hardware. Redundancy makes it is easy to see how fairly sharp dissociations could arise—knocking out one token module need

disrupt only one high-level operation: other high-level operations that draw on the same *type* of resource may well be spared.

The consequences of this distinction between literal reuse and reuse by proxy for much speculation about the localization and specialization of function are potentially profound. In cognitive neuropsychology the discovery that a focal lesion selectively impairs a particular cognitive function is routinely taken as evidence of its functional specificity (Coltheart 2011; Sternberg 2011). Even cognitive scientists who take a developmental approach to modularity, i.e. who concede that parts of the mind may be modular but stress that modularization is a developmental process, concede too much when they imply, as they frequently do, that modularization results in domain-specific modules (Karmiloff-Smith 1992; Prinz 2006; Barrett 2006; Cowie 2008; Guida et al. 2016). This is true in some sense, but not in anything like the standard sense, for redundancy envisages that developmental modules form a special class of neural networks, namely those which are *qualitatively* identical but *numerically* distinct. The appearance of modularization in development is thus fully compatible with deep domain interpenetration. In any event redundancy does not predict that all acquired skills will be modular. The evidence suggests that while some complex skills reside in at least partly dissociable circuitry, most complex skills are implemented in more typical neural networks, i.e. those consisting of literally shared parts.[8]

[8] This seems to be true regardless of whether the complex skills are innate or acquired.

4. What Else Might Redundancy Explain?

It is generally a good design feature of any system to have spare capacity. For instance, in engineered systems, "redundant parts can substitute for others that malfunction or fail, or augment output when demand for a particular output increases" (Whiteacre 2010, 14). The positive connection between robustness and redundancy in biological systems is also clear (Edelman and Gally 2001; Mason 2010; Whiteacre 2010; Iriki & Taoka 2012). So there are good reasons for evolution to have seen to it that our brains have spare capacity. But in the case of the brain and the cortex most especially, there are other reasons why redundancy would be an important design feature. It offers a solution to what Jungé and Dennett (2010, 278) called the "time-sharing" problem. It may also offer a solution to what I call the "encapsulation" problem.

The time-sharing problem arises when multiple simultaneous demands are made on the same cognitive resource. This is probably a regular occurrence. Here are just a few examples.

- Driving a car and holding a conversation at the same time: if it is true that some of the selfsame motor operations underlying aspects of speech production and comprehension are also required for the execution of sequenced or complex motor functions (Pulvermüller and Fadiga 2010; Graziano et al. 2002; MacNeilage 1998; Glenberg et al.

2008; Glenberg and Kaschak 2002; Glenberg et al. 2007; Greenfield 1991), as perhaps exemplified by driving a manual vehicle or operating complex machinery (e.g. playing the organ), how do we manage to pull this off?

- By reflecting the recursive structure of thought (Christiansen and Chater 2016, 51), the language circuits may redeploy a recursive operation simultaneously during sentence production. This might be the case during the formation of an embedded relative clause—the thought and its encoding may require parallel use of the same sequencing principle. Again, how do we manage this feat?
- If metarepresentational operations are involved in the internalization of conventional sound-meaning pairs, and also in the pragmatics and mindreading that carry on simultaneously during conversation, as argued by Suddendorf (2013), could this not simply be another instance of time-sharing? The example is contentious, but it still raises the question: how does our brain manage to do things like this?
- Christiansen and Chater's (2016) "Chunk and Pass" model of language processing envisages *multilevel* and *simultaneous* chunking procedures. As they put it, "the challenge of language acquisition is to learn a dazzling sequence of rapid processing operations" (2016, 116). What must the brain be like to allow for this dazzling display?

Explaining these phenomena is difficult. Indeed when dealing with clear (literal) instances of reuse, results from the interference paradigm show that processing bottlenecks are inevitable—true multi-tasking is impossible. Redundancy offers a natural explanation of how

Seattle, WA; 1-4 November 2018 -543-

the brain overcomes the time-sharing problem. It explains, in short, how we are able to walk and chew gum at the same time.

Redundancy might also offer a solution to what I have called the encapsulation problem. The neural networks that implement cognitive functions are not likely to be characterized by informational encapsulation if they share their nodes with networks implementing other cognitive functions. This is because in sharing their nodes with these other systems they will prima facie have access to the information stored and manipulated by those other systems (Anderson 2010, 300). If, then, overlapping brain networks must share information (Pessoa 2016, 23), it would be reasonable to suppose that central and peripheral systems do *not* overlap. For peripheral systems, which are paradigmatically fast and automatic, would not be able to process inputs as efficiently if there were a serious risk of central system override—i.e. of beliefs and other central information getting in the way of automatic processing. But we know from the neuroimaging literature that quite often the brain networks implementing central and peripheral functions *do* overlap. This is puzzling in light of the degree of cognitive impenetrability that certain sensory systems still seem to exhibit—limited though it may be. If it is plausible to suppose that the phenomenon calls for segregated circuitry, redundancy could feature in a solution to the puzzle, since it naturally explains how the brain can make parallel use of the same resources. Neuroimaging maps might well display what appear to be overlapping brain regions between two tasks (one involving central information, the other involving classically peripheral operations), but the

overlap would not exist—there would be distinct albeit adjacent or interdigitated and nearly identical circuits recruited in each case. Of course there may be other ways around the encapsulation problem that do not require segregated circuitry: the nature and extent of the overlap is presumably important. But clearly redundancy opens up some fascinating explanatory possibilities.

To the extent that acquired skills must overcome both the time-sharing problem as well as the encapsulation problem—for acquired competencies are often able to run autonomously of central processes—we might expect that their neural implementations incorporate redundant tissue. In concluding, let me illustrate this point by offering a gloss on a particular account of how skills and expertise are acquired during development elaborated by Guida et al. (2016) and Anderson (2014). The process involved is called "search" (Anderson 2014). Search is an exploratory synaptogenetic process, "the active testing of multiple neuronal combinations until finding the most appropriate one for a specific skill, i.e., the neural niche of that skill" (Guido et al. 2016, 13). The theory holds that in the early stages of skill acquisition, the brain must search for an appropriate mix of brain areas, and does so by recruiting relatively widely across the cortex. When expertise has finally developed, a much narrower and more specific network of brain areas has been settled upon, such that "[a]s a consequence of their extended practice, experts develop domain-specific knowledge structures" (Guido et al. 2016, 13). The gloss (and my hunch) is this: first, that repeated practice of a task that requires segregation (to get around time-

sharing and encapsulation issues) will in effect *force* search into redundant neural territory (Karmiloff-Smith 1992; Barrett 2006; Barret and Kurzban 2006); second, that search will recruit idle or relatively underutilized circuits in preference to busy ones as a general default strategy. Guido et al. (2016) cite evidence that experts' brains reuse areas for which novices' brains make only limited use: "Whereas novices use episodic long-term memory areas (e.g., the mediotemporal lobe) for performing long-term memory tasks, experts are able to (re)use these areas also for performing working-memory tasks" (Guido et al. 2016, 14). Guido and colleagues, in agreement with Anderson (2014), seem to have literal reuse in mind. But the same evidence they cite is consistent with reuse by proxy. As Barrett and Kurzban (2006, 639) suggest, echoing a similar suggestion by Karmiloff-Smith (1992), a developmental system

> could contain a procedure or mechanism that partitioned off certain tasks—shunting them into a dedicated developmental pathway—under certain conditions, for example, when the cue structure of repeated instances of the task clustered tightly together, and when it was encountered repeatedly, as when highly practiced....Under this scenario, reading could still be recruiting an evolved system for object recognition, and yet phenotypically there could be *distinct modules* for reading and for other types of object recognition.

5. Conclusion

It is true that language and other cognitive skills frequently dissociate from other skills, but redundancy puts this sort of modularization in its proper context. Redundancy predicates functional inheritance across tasks and task categories even when the tasks are implemented in spatially segregated neural networks. Thus dissociation evidence alone does not always indicate true functional specificity. In particular, these dissociations provide no evidence that language is cognitively special vis-à-vis other cognitive domains.

References

Anderson, Michael L. 2007a. "Evolution of Cognitive Function via Redeployment of Brain Areas." *The Neuroscientist* 13:13-21.

——2007b. "Massive Redeployment, Exaptation, and the Functional Integration of Cognitive Operations." *Synthese* 159 (3): 329-345.

——2007c. "The Massive Redeployment Hypothesis and the Functional Topography of the Brain." *Philosophical Psychology* 21 (2): 143-174.

——2008. "Circuit Sharing and the Implementation of Intelligent Systems." *Connection Science* 20 (4): 239-251.

——2010. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences* 33 (4): 245-266; discussion 266-313.

——2014. *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Anderson, Michael L., and Barbara L. Finlay. 2014. "Allocating Structure to Function: The Strong Links Between Neuroplasticity and Natural Selection." *Frontiers in Human Neuroscience* 7:1-16.

Apperly, I.A., D. Samson, N. Carroll, S. Hussain, and G. Humphreys. 2006. "Intact First- and Second-Order False Belief Reasoning in a Patient with Severely Impaired Grammar." *Social Neuroscience* 1 (3-4): 334-348.

Barrett, H. Clark. 2006. "Modularity and Design Reincarnation." In *The Innate Mind Volume 2: Culture and Cognition*, ed. Peter Carruthers, Stephen Laurence, and Stephen P. Stich, 199-217. New York: Oxford University Press.

Barrett, H. Clark, and Robert Kurzban. 2006. "Modularity in Cognition: Framing the Debate." *Psychological Review* 113 (3): 628-647.

Bergeron, Vincent. 2007. "Anatomical and Functional Modularity in Cognitive Science: Shifting the Focus." *Philosophical Psychology* 20 (2): 175-195.

Buxhoeveden, Daniel P., and Manuel F. Casanova. 2002. "The Minicolumn Hypothesis in Neuroscience." *Brain* 125:935-951.

Carruthers, Peter. 2006. The *Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Seattle, WA; 1-4 November 2018 -549-

Casasanto, D., and K. Dijkstra. 2010. "Motor Action and Emotional Memory." *Cognition* 115 (1): 179-185.

Chomsky, Noam. 1980. *Rules and Representations*. New York: Columbia University Press.

——1988. *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.

——2002. *On Nature and Language*. New York: Cambridge University Press.

Christiansen, Morten H., and Nick Chater. 2016. *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge, MA: MIT Press.

Coltheart, Max. 1999. "Modularity and Cognition." *Trends in Cognitive Sciences* 3 (3): 115-120.

——2011. "Methods for Modular Modelling: Additive Factors and Cognitive Neuropsychology." *Cognitive Neuropsychology* 28 (3-4): 224-240.

Cosmides, Leda, and John Tooby. 1994. "Origins of Domain Specificity: The Evolution of Functional Organization." In *Mapping the World: Domain Specificity in Cognition and Culture*, ed. L. Hirschfield, and S. Gelman, 85-116. New York: Cambridge University Press.

Cowie, Fiona. 2008. "Innateness and Language." In *The Stanford Encyclopedia of Philosophy*, winter 2016, ed. E.N. Zalta. <http://plato.stanford.edu/archives/win2016/entries/innateness-language/>

de Almeida, Roberto G., and Lila R. Gleitman, eds. 2018. *On Concepts, Modules, and Language: Cognitive Science at its Core*. New York: Oxford University Press.

Deacon, Terrence W. 2010. "A Role for Relaxed Selection in the Evolution of the Language Capacity." *Proceedings of the National Academy of Sciences of the United States of America* 107: 9000-9006.

Dehaene, Stanislas. 2005. "Evolution of Human Cortical Circuits for Reading and Arithmetic: The 'Neuronal Recycling' Hypothesis." In *From Monkey Brain to Human Brain*, eds. Stanislas Dehaene, J.R. Duhamel, M.D. Hauser, and G. Rizzolatti, 133-157. Cambridge, MA: MIT Press.

Edelman, Gerald M., and Joseph A. Gally. 2001. "Degeneracy and Complexity in Biological Systems." *Proceedings of the National Academy of Sciences of the United States of America* 98 (24): 13763-13768.

Eliasmith, Chris. 2015. "Building a Behaving Brain." In *The Future of the Brain*, ed. Gary Marcus, and Jeremy Freeman, 125-136. Princeton: Princeton University Press.

Fedorenko, Evelina, Michael K. Behr, and Nancy Kanwisher. 2011. "Functional Specificity for High-Level Linguistic Processing in the Human Brain." *Proceedings of the National Academy of Sciences of the United States of America* 108 (39): 16428-16433.

Fedorenko, Evelina, John Duncan, and Nancy Kanwisher. 2012. "Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area." *Current Biology* 22 (21): 2059-2062.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Friston, Karl J., and Cathy J. Price. 2003. "Degeneracy and Redundancy in Cognitive Anatomy." *Trends in Cognitive Sciences* 7 (4): 151-152.

Gauthier, I., T. Curran, K.M. Curby, and D. Collins. 2003. "Perceptual Interference Supports a Non-Modular Account of Face Processing." *Nature Neuroscience* 6 (4): 428-432.

Glenberg, A.M., M. Brown, and J.R. Levin. 2007. "Enhancing Comprehension in Small Reading Groups Using a Manipulation Strategy." *Contemporary Educational Psychology* 32:389-399.

Glenberg, A.M., and M.P. Kaschak. 2002. "Grounding Language in Action." *Psychonomic Bulletin and Review* 9:558-565.

Glenberg, A.M., M. Sato, and L. Cattaneo. 2008. "Use-Induced Motor Plasticity Affects the Processing of Abstract and Concrete Language." *Current Biology* 18 (7): R290-291.

Godfrey-Smith, Peter. 2001. "Three Kinds of Adaptationism." In *Adaptationism and Optimality*, ed. Steven H. Orzack, and Elliott Sober, 335-357. Cambridge: Cambridge University Press.

Graziano, M.S.A., C.S.R. Taylor, T. Moore, and D.F. Cooke. 2002. "The Cortical Control of Movement Revisited." *Neuron* 36:349-362.

Greenfield, P.M. 1991. "Language, Tools and Brain: The Ontogeny and Phylogeny of Hierarchically Organized Sequential Behavior." *Behavioral and Brain Sciences* 14 (4): 531- 551; discussion 551-595.

Guida, Alessandro, Guillermo Campitelli, and Fernand Gobet. 2016. "Becoming an Expert: Ontogeny of Expertise as an Example of Neural Reuse." *Behavioral and Brain Sciences* 39:13-15.

Hickok, G., and David Poeppel. 2000. "Towards a functional neuroanatomy of speech perception." *Trends in Cognitive Sciences* 4 (4): 131-138.

Iriki, Atsushi, and Miki Taoka. 2012. "Triadic (ecological, neural, cognitive) niche construction: A scenario of human brain evolution extrapolating tool use and language from the control of reaching actions." *Philosophical Transactions of the Royal Society B* 367: 10-23.

Jungé, Justin A., and Daniel C. Dennett. 2010. "Multi-Use and Constraints from Original Use." *Behavioral and Brain Sciences* 33 (4): 277-278.

Kandel, E.R., and A.J. Hudspeth. 2013. "The Brain and Behavior." In *Principles of Neural Science*, ed. E.R. Kandel, J.H. Schwartz, T.M. Jessell, S.A. Siegelbaum, and A.J. Hudspeth, 5-20. New York: McGraw-Hill.

Karmiloff-Smith, Annette. 1992. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Leo, Andrea, Giulio Bernardi, Giacomo Handjaras, Daniela Bonino, Emiliano Ricciardi, and Pietro Pietrini. 2012. "Increased BOLD Variability in the Parietal Cortex and Enhanced Parieto-Occipital Connectivity During Tactile Perception in Congenitally Blind Individuals." *Neural Plasticity* 2012:1-8 doi: 10.1155/2012/720278.

Luria, A.R., L.S. Tsvetkova, and D.S. Futer. 1965. "Aphasia in a Composer (V.G. Shebalin)." *Journal of the Neurological Sciences* 2 (3): 288-292.

MacNeilage, P.F. 1998. "The Frame/Content Theory of Evolution of Speech Production." *Behavioral and Brain Sciences* 21 (4): 499-511; discussion 511-546.

Maleszka, Ryszard, Paul H. Mason, and Andrew B. Barron. 2013. "Epigenomics and the Concept of Degeneracy in Biological Systems." *Briefings in Functional Genomics* 13 (3): 191-202.

Mason, Paul H. 2010. "Degeneracy at Multiple Levels of Complexity." *Biological Theory* 5 (3): 277-288.

McGeer, Victoria. 2007. "Why Neuroscience Matters to Cognitive Neuropsychology." *Synthese* 159:347-371.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Pessoa, Luiz. 2016. "Beyond Disjoint Brain Networks: Overlapping Networks for Cognition and Emotion." *Behavioral and Brain Sciences* 39:22-24.

Peretz, Isabelle., and Max Coltheart. 2003. "Modularity of music processing." *Nature Neuroscience* 6:688-691.

Pinker, Steven, and Ray Jackendoff. 2005. "The Faculty of Language: What's Special About It?" *Cognition* 95:201-236.

Plaut, David C. 1995. "Double Dissociation Without Modularity: Evidence from Connectionist Neuropsychology." *Journal of Clinical and Experimental Psychology* 17 (2): 291-321.

Poeppel, David. 2001. "Pure Word Deafness and the Bilateral Processing of the Speech Code." *Cognitive Science* 21 (5): 679-693.

Prinz, Jesse J. 2006. "Is the Mind Really Modular?" In *Contemporary Debates in Cognitive Science*, ed. R. Stainton, 22-36. Oxford: Blackwell.

Pulvermüller, Friedmann, and Luciano Fadiga. 2010. "Active Perception: Sensorimotor Circuits as a Cortical Basis for Language." N*ature Reviews Neuroscience* 11:351-360.

Sternberg, Saul. 2011. "Modular Processes in Mind and Brain." *Cognitive Neuropsychology* 28 (3-4): 156-208.

Striem-Amit, Ella, and Amir Amedi. 2014. "Visual Cortex Extrastriate Body-Selective Area Activation in Congenitally Blind People 'Seeing' by Using Sounds." *Current Biology* 24:1-6.

Suddendorf, Thomas. 2013. *The Gap: The Science of What Separates Us from the Animals*. New York: Basic Books.

Varley, R.A., N.J.C. Klessinger, C.A.J. Romanowski, and M. Siegal. 2005. "Agrammatic But Numerate." *Proceedings of the National Academy of Sciences of the United States of America* 102:3519-3524.

Whiteacre, James M. 2010. "Degeneracy: A Link Between Evolvability, Robustness and Complexity in Biological Systems." *Theoretical Biology and Medical Modelling* 7 (6): 1-17.

Zerilli, John. 2017a. "Against the 'System' Module." *Philosophical Psychology* 30 (3): 235-250.

——2017b. "Multiple Realization and the Commensurability of Taxonomies." *Synthese* (https://doi.org/10.1007/s11229-017-1599-1).

Dissolving the Measurement Problem Is Not an Option for the Realist

February 28, 2018

This paper critically assesses the proposal that scientific realists do not need to search for a solution of the measurement problem in quantum mechanics, but should instead dismiss the problem as ill-posed. James Ladyman and Don Ross have sought to support this proposal with arguments drawn from ontic structural realism and from a Bohr-inspired approach to quantum mechanics. I show that the first class of arguments is unsuccessful, because formulating the measurement problem does not depend on the metaphysical commitments which are undermined by ontic structural realism. The second class of arguments is problematic due to its refusal to provide an analysis of the term 'measurement". It turns out that the proposed dissolution of the measurement problem is in conflict not only with traditional forms of scientific realism but even with the rather minimal realism that Ladyman and Ross themselves defend.

1 Introduction

One of the attractions of non-realist approaches to quantum mechanics (QM) is that they offer a way to dissolve the measurement problem instead of adopting a solution for it that would either have to modify the physics (such as theories with additional variables or spontaneous collapses) or drastically inflate the empirically inaccessible content of reality (such as many-worlds interpretations). Nonetheless, many metaphysicians (and some physicists) consider the abandonment of realism too high a price to pay and therefore insist that the measurement problem calls for a (realistic) solution rather than a dissolution along non-realist lines.

Could there be a third position that somehow combines the attractions of these two (seemingly opposing) camps? In this paper, I critically assess a recent proposal for such a position, which can be found in the writings of James Ladyman and Don Ross (2007; 2013). While describing their approach to the measurement problem as "roughly the sort of account favoured by earlier versions of the Copenhagen interpretation" (2013, 134), Ladyman and Ross do not think this amounts to instrumentalism, but rather seek

to defend it as part of their brand of scientific realism, which combines ontic structural realism (OSR) with what they call "rainforest realism" (2007, Chapter 4). The purpose of the present paper is to show that their Copenhagen-style dissolution of the measurement problem does not fit well with a position that presents itself as a version of scientific realism.

In order to avoid misunderstandings, let me first mention an important point of agreement between Ladyman/Ross and myself. Much of their (2013) argument is directed against the simple realism-versus-instrumentalism dichotomy (with respect to QM), which they find operative in the philosophy of David Deutsch (2011). More specifically, they argue that there is much within the formalism of QM about which one can be a realist despite not being committed to any realistic solution of the measurement problem. I largely agree with this claim, noting that one need not be an ontic structural realist to appreciate this point (cf. Cordero 2001; Saatsi forthcoming). Nor am I particularly worried about Michael Esfeld's (2013) diagnosis of OSR being only a partial realism if it does not incorporate a realist treatment of the measurement problem. What I do criticize is that the dissolution of the measurement problem proposed by Ladyman and Ross undermines some specific commitments that should be part of any position deserving to be called realism (even only a partial one).

My investigation proceeds as follows. In Section 2, I will consider how OSR might seem to undermine the traditional formulation of the measurement problem, insofar as the latter depends on viewing measurement devices as being composed of quantum particles. My contention will be that while OSR incorporates some telling arguments against traditional accounts of composition, none of them justifies viewing the measurement problem as a pseudo-problem. Section 3 will then show how Ladyman's and Ross's Bohr-inspired alternative approach to the measurement problem contradicts some widely shared realist (and, relatedly, naturalist) commitments. This by itself may not be so problematic for Ladyman and Ross, because they do not subscribe to standard realism anyway. I will therefore complete my argument in Section 4 by demonstrating that some of Ladyman's and Ross's own arguments in favor of scientific realism are in tension with their proposed dissolution of the measurement problem.

2 Composition and the measurement problem

A key step in setting up the measurement problem consists in assigning a quantum state to a measurement apparatus. In the standard example (see, e.g., Myrvold 2017, Section 4), two possible final states of the apparatus are denoted by $|\text{“0”}\rangle_A$ and $|\text{“1”}\rangle_A$, corresponding to the two basis states $|0\rangle_S$ and $|1\rangle_S$ of the measured quantum system. The measurement problem then consists in making sense of the superposed final state $a|0\rangle_S|\text{“0”}\rangle_A + b|1\rangle_S|\text{“1”}\rangle_A$, which, if a and b are both nonzero, does not seem to describe anything we observe.

Obviously, this whole construction crumbles if one refuses to assign quantum states to measurement devices. Why should the quantum formalism be the appropriate tool for describing such commonsensical objects? The usual rationale is that these objects are

assumed to be composed of (a very large number of) quantum particles, and this is precisely the assumption that OSR rejects. Let us therefore look whether the arguments for this rejection may also serve to undermine the formulation of the measurement problem just given.

The hostility to the idea that material bodies are made of little things is one of the recurring themes in the work of Ladyman and Ross. Early on in their seminal (2007) book *Every Thing Must Go*, they describe how analytic metaphysicians have been misled by their hankering after a general *a priori* notion of composition, which pays little or no attention to what we have learnt about specific composition relations described within various sciences. In conclusion, they note: "We have no reason to believe that an abstract composition relation is anything other than an entrenched philosophical fetish" (21).

However, this kind of criticism can be dealt with rather quickly for our purpose, by noting that no abstract metaphysical composition relation is presupposed by the claim that measurement devices are composed of quantum systems. Instead, QM itself furnishes us with the rules of how systems combine to form larger systems, and it is those rules that tell us that the compound systems can be in superposed states just as the elementary systems can. Furthermore, insofar as the quantum formalism describes interactions between the parts of compound systems, the quantum mechanical composition relation also has the necessary dynamical character that Ladyman (2017, 156) finds missing in the traditional metaphysical accounts of composition.

But of course, the mere fact that QM allows us to describe systems composed of many particles does not justify the idea that measurement devices ought to be so described, or even that they ought to be described quantum mechanically at all. In the words of Ladyman and Ross (2007, 182), "the application of the quantum formalism to macroscopic objects is not necessarily justified, especially if those objects are importantly different from microscopic objects, as indeed they are, in not being carefully isolated from the environment".

The suggestion that a quantum description becomes inappropriate to the extent that systems fail to be isolated has some initial plausibility, as the interference effects indicating quantum behavior are indeed only observed for systems that are sufficiently well isolated from their environment. In that sense, application of the quantum formalism to macroscopic objects lacks direct empirical justification. It would, however, be hasty to infer from this lack of empirical justification a lack of scientific justification in a more general sense. To see why, we should note that the most important recent progress in the study of how the lack of isolation influences the behavior of physical systems has been through the theory of decoherence (see Bacciagaluppi 2016 for a review). By operating entirely within the formalism of QM, this theory presupposes just what Ladyman and Ross seek to prohibit: the assignment of quantum states to macroscopic objects (the environment).[1] Such assignments are therefore not just a philosopher's fancy, but are soundly rooted in scientific theorizing.

A further possible reason to reject the idea of macroscopic objects being composed

[1] As is well known, the fact that decoherence theory is just an application of standard QM also implies that decoherence by itself does not solve the measurement problem (Bacciagaluppi 2016, Section 2.1).

of quantum particles is that, according to OSR, there are no particles. As the term "particle" can have different meanings, not all of which may be relevant for our purpose, we need to look at the different anti-particle arguments advanced by OSR and evaluate for each of them whether or not it threatens the usual way of setting up the measurement problem.

The historically most important and most widely discussed issue in OSR's account of particles concerns their identity and individuality (or lack thereof; see Ladyman 2016 and references therein). This issue, however, does not seem to have much relevance for the pertinence of viewing macroscopic objects as composed of quantum systems. On the contrary, the fact that there is a metaphysical debate on the (non-)individuality of quantum particles at all precisely shows that the quantum mechanics of composite systems is to some extent insensitive to whether the components are regarded as individuals or not. It is only their cardinality that matters, and this latter is unproblematic in non-relativistic QM (I will turn to relativistic quantum theory in a minute). Another way to make the same point is to note that the formalism of quantum mechanics already incorporates the features that fueled the debate on identity and individuality (namely, the indistinguishability postulate in quantum statistics and the non-supervenience of entanglement relations), so one should not expect the results of that debate to undermine the quantum mechanical account of composition.

Neither is it relevant whether particles are regarded as elementary (or fundamental) in any strong sense. I fully agree with Ladyman (2016, 202) that we should not regard them in this way, but nothing in the usual formulation of the measurement problem depends on doing so. In fact, there is excellent empirical evidence for the occurrence of superposed states in unambiguously non-elementary quantum systems (see Arndt and Hornberger 2014 for a recent review), so the whole problem can be set up without any reference to elementary or fundamental particles.

At this point, we should attend to the fact that Ladyman and Ross are not only committed to OSR but also to rainforest realism, which furnishes a sense in which even they agree that particles do exist: they are real patterns (Ladyman 2016, 203-204). This is all the more important as the particle concept becomes increasingly problematic when we turn from non-relativistic quantum mechanics to relativistic quantum field theory. The appearance of particle creation and annihilation not only undermines the above-mentioned appeal to a well-defined cardinality of particles in a composite system, but it also blurs the line between particles as persisting objects and mere excitations of quantum fields. More precisely, this latter distinction now becomes dependent on the time and energy scale at which a system is considered (Ladyman 2017, 158). The merit of rainforest realism is that it takes this dependence into account and makes room for scale-relative ontological commitments.

This implies that reference to particles is unproblematic as long as the context is appropriately specified in terms of the relevant time and energy scale. Formulations of the measurement problem implicitly do this by involving only two kinds of physical systems, namely non-relativistic quantum particles and macroscopic measurement devices. Neither of them requires consideration of quantum field theoretic effects, hence the scale at which the measurement problem is formulated is not affected by the breakdown of the

particle concept in quantum field theory.

In the same context, Ladyman (2017, 156-157) mentions the renormalization group to illustrate the vast difference between the naïve metaphysical picture of composition and the intricate way in which actual condensed matter physics describes how gross matter behaves in terms of interactions between atoms, electrons, and fields. Again, the reference to scientific accounts of composition is well taken, but I do not see how it could undermine the idea that matter is composed of quantum particles. In an important sense, this latter idea provides the very motivation for applying renormalization group methods in condensed matter physics, as a tool to eliminate degrees of freedom associated with the atomic constitution of matter that are irrelevant for its macroscopic behavior. At the same time, quantum measurement devices are characterized by the fact that *some* microscopic degrees of freedom (namely the ones being measured) are *not* eliminated, but do indeed affect the device's macroscopic behavior. The renormalization group has nothing to say about these degrees of freedom and it therefore does not tell against describing the measurement device with respect to them in the way that gives rise to the measurement problem.

3 Leaving "measurement" unanalyzed?

The previous section has shown that OSR's rejection of naïve views about the constitution of matter does not justify dismissal of the measurement problem as a pseudo-problem. But Ladyman and Ross (2007, 182) also propose a more general reason for dismissal, based on their commitment to naturalism:

> From the point of view of the [Principle of Naturalistic Closure], the representation of macroscopic objects using quantum states can only be justified on the basis of its explanatory and predictive power and it has neither. ... The predictive success of QM in this context consists in the successful application of the Born rule, and that is bought at the cost of a pragmatic splitting of the world into system and apparatus.

I have already noted above (with reference to decoherence theory) that the naturalistic credentials of assigning quantum states to macroscopic objects may be better than Ladyman and Ross suppose, so let me now focus on the positive part of their proposal. The application of the Born rule is indeed successful if we simply insist (as Niels Bohr famously did) that the apparatus needs to be described classically in the sense of not being in any superposed state. Ladyman and Ross (2013, 134) explicitly sympathize with Bohr's early version of the Copenhagen interpretation, which they view as distinct from later versions in virtue of its refusal to give any story about collapse of the wave function.

Without entering into the complex debate on the history of the Copenhagen interpretation, it is noteworthy that Ladyman and Ross (ibid.) identify "an abandonment not so much of realism as of naturalism itself" in the transition from Bohr to later versions

of Copenhagen, which "*did* include a story about collapse, but interpreted it as a consequence of measurement". Against this assessment, I submit that the abandonment of naturalism (and realism) takes place when one endorses Bohr's version of Copenhagen (more precisely: Ladyman's and Ross's reading of it), not when one switches from Bohr to a later version.

Ladyman's and Ross's argument for viewing Bohr's approach as compatible with scientific realism depends on the interpretation-independent content of standard QM already mentioned in Section 1. However, as we just saw, some of that content gets its empirical character only via the successful application of the Born rule — the content is interpretation-independent precisely because any viable interpretation of QM needs to incorporate the Born rule in some way.

Now the problem with the Born rule is that it speaks about the probabilities of measurement results, while it is notoriously unclear what counts as a "measurement". This critique is well known, and it is often put in terms of awkward questions for those who (in the spirit of later Copenhagen) tie the notion of measurement to a collapse of the wave function. So for example, John S. Bell (1990, 19) famously asked whether a single-celled living creature already qualified to play the role of "measurer" or whether it takes some better qualified system (with a Ph.D.?) to make the wave function collapse. But the basic point of criticism (as Bell makes clear in the rest of his paper) does not depend on any specific view of wave function collapse, but on using such a desperately imprecise notion as "measurement" in a basic assumption of physics. This is why realistic versions of QM (such as those associated with the names of Everett, Bohm, or GRW) seek to *derive* the Born rule by giving a physical account of what it is to be a measurement. Bohr, on the other hand, denies the need for such an account, as Ladyman and Ross (2013, 134) point out approvingly.

Admittedly, any theory has to operate with some basic notions which are not amenable to further analysis, so why not simply treat "measurement" as such a notion? This works well for situations in which we all agree whether the notion applies or not. But what about ambiguous cases, for example, a device that displays a measurement outcome which is not (even indirectly) observed[2] by anyone? Bohr repeatedly insisted that human observers play no essential role in the measurement process, but how can this be justified without an analysis of "measurement"? Neither our pre-theoretical nor our scientific usage of the word "measurement" seems to settle the question whether unobserved measurements should still count as measurements.

A verificationist will denounce this as yet another pseudo-question, because such events (by definition) do not make any difference to what we observe, hence we should not suppose that there are any matters of fact concerning them. But this is hard to square with realism, understood as a stance that refuses to limit reality to what we can observe, or worse still, to what we actually *do* observe. Ladyman and Ross (2007, 309) are quite honest about how their verificationism limits the domain of what counts as real, but the

[2] A useful explication of the notion of "observation" relevant for this context is given by Ladyman and Ross (2007, 307) in terms of "informational connectedness". In the following, I have this rather wide sense of "observation" in mind when I speak of "unobserved measurements" or "unobserved data".

conflict with realism is obscured by the somewhat far-fetched example they give in that context: Most realists will readily agree that "there are no grounds for regarding the other side of [the Big Bang] as part of reality" (ibid.). By contrast, many realists will think that something has gone deeply wrong if we are discouraged from believing that there is a fact of the matter as to how our measurement devices behave when no one watches them.

Before I turn (in the next section) to the question in how far Ladyman's and Ross's own version of realism should be bothered by this tension, I should also mention that there is something anti-naturalistic about drawing such anthropocentric limits around what counts as real. A thorough discussion of Ladyman's and Ross's (2007, Section 6.3) arguments for the compatibility of naturalism and verificationism is beyond the scope of this paper. Suffice it to say that these arguments are most plausible when the verifiability criterion is understood epistemically (as a policy on what our theorizing should or should not be concerned with) rather than metaphysically (as a criterion on what does or does not belong to reality). To the extent that the latter reading is implied, a naturalist is likely to wonder why reality should care which parts of it are accessible to our observation.

4 Unobserved measurements and objective modality

That the Bohrian approach to the measurement problem entails a conflict with some widespread realistic and naturalistic intuitions may not be a decisive reason against it, especially if one has already abandoned certain commitments of standard realism, as proponents of OSR have. I will therefore now try to show that the proposed dissolution of the measurement problem conflicts not only with standard realism but also with elements of scientific realism that Ladyman and Ross themselves endorse.

A first hint of this conflict appears in the role that the notion of "data" plays within rainforest realism. As we saw in Section 2, Ladyman and Ross conceive of reality in terms of real patterns, and patterns are "relations among data" (2007, 228). In their discussion of Dennett's (1991) account of real patterns, Ladyman and Ross carefully distance themselves from the kind of instrumentalism that is at least partly invited by Dennett's writing and has preoccupied many of his commentators. In the process of doing so, they acknowledge that "there are (presumably) real patterns in lifeless parts of the universe that no actual observer will ever reach" (Ladyman and Ross 2007, 203). But such realism about patterns presupposes realism about data regardless of whether they are observed or not. It is therefore in tension with non-realism about unobserved quantum measurements.

The problem comes into sharper focus when we turn to Ladyman's and Ross's (2007, Subsection 2.3.2) critique of van Fraassen's constructive empiricism. While they largely share van Fraassen's aversion to traditional metaphysics, they defend a commitment to objective modality as a crucial element of realism against his deflationary view. One reason for this is that "theories are always modalized in the sense that they allow for a variety of different initial conditions or background assumptions rather than just the actual ones, and so describe counterfactual states of affairs" (110). A constructive em-

piricist might regard the claim that science gives us knowledge about non-actual states of affairs as unjustified, because all we ever experience is the actual. But this, according to Ladyman and Ross, neglects the fact that we can to some extent vary what becomes actual and still experience that our theories accurately predict what we observe. In other words, the empiricist relies on a somewhat arbitrary boundary when confining the content of our theories to a description of what actually occurs.

Insofar as this accurately describes the motivation for preferring OSR to constructive empiricism, an adherent of OSR should be equally dissatisfied with versions of QM which fail to give a non-anthropocentric account of measurement, because they involve a similar boundary between what our theories do and do not tell us. In this case, it is not the boundary between what actually occurred and what could have occurred under different initial conditions (if the Born rule is modalized in the above sense, it does give us knowledge about both of these), but the boundary between what was actually observed and what actually occurred without being observed (the Born rule being silent about the latter set of events). This second boundary is just as arbitrary as the first one, because it is largely up to us which occurrent events are observed and which ones are not.

The same point can also be made in terms of demands for explanation. In general, Ladyman and Ross share van Fraassen's s skepticism towards such demands, but here is one they explicitly accept: "That we are so often able to identify regularities in phenomena and then use them for prediction needs to be explained" (Ladyman and Ross 2007, 106). If OSR is to have any advantage over constructive empiricism, the sought-after explanation cannot simply be that there are such regularities, because that explanation would be available to the constructive empiricist as well. In order to satisfy OSR's demand, the regularities need to be invested with modal force, which enables us to answer questions about counterfactual situations. Among such questions are those about what would have happened if we had not been around to observe the phenomena in question, and an explanation would hardly be deemed satisfactory if it postulated regularities that only obtain if some observer is present. But this is precisely what the Born rule does, if it is interpreted as a modally charged law but not supplemented by a non-anthropocentric account of "measurement".

To sum up, the proposal to dissolve the measurement problem along Bohrian lines conflicts not only with some commitments of standard realism (as demonstrated in Section 3) but also with the rather minimal kind of realism that Ladyman and Ross defend against van Fraassen. Furthermore, Section 2 has shown that none of the arguments for OSR serve to undermine the view that measurement devices are composed of quantum systems in the sense relevant for formulating the measurement problem. Therefore, scientific realists (including proponents of OSR) should acknowledge that the measurement problem calls for a solution, not a mere dissolution.

References

Arndt, M. and K. Hornberger (2014). Testing the limits of quantum mechanical superpositions. *Nature Physics 10*, 271–277.

Bacciagaluppi, G. (2016). The role of decoherence in quantum mechanics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2016 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2016/entries/qm-decoherence/`.

Bell, J. S. (1990). Against 'measurement'. In A. I. Miller (Ed.), *Sixty-two years of uncertainty: historical, philosophical, and physical inquiries into the foundations of quantum mechanics*, pp. 17–32. New York: Plenum Press.

Cordero, A. (2001). Realism and underdetermination: Some clues from the practices-up. *Philosophy of Science 68*, S301–S312.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy 88*, 27–51.

Deutsch, D. (2011). *The Beginning of Infinity*. London: Allen Lane.

Esfeld, M. (2013). Ontic structural realism and the interpretation of quantum mechanics. *European Journal for Philosophy of Science 3*, 19–32.

Ladyman, J. (2016). Are there individuals in physics, and if so, what are they? In A. Guay and T. Pradeu (Eds.), *Individuals Across the Sciences*, pp. 193–206. Oxford: Oxford University Press.

Ladyman, J. (2017). An apology for naturalized metaphysics. In M. H. Slater and Z. Yudell (Eds.), *Metaphysics and the Philosophy of Science: New Essays*, pp. 141–161. New York: Oxford University Press.

Ladyman, J. and D. Ross (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Ladyman, J. and D. Ross (2013). The world in the data. In D. Ross, J. Ladyman, and H. Kincaid (Eds.), *Scientific Metaphysics*, pp. 108–150. Oxford: Oxford University Press.

Myrvold, W. (2017). Philosophical issues in quantum theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2017/entries/qt-issues/`.

Saatsi, J. (forthcoming). Scientific realism meets metaphysics of quantum mechanics. In A. Cordero (Ed.), *Philosophers Think About Quantum Theory*. Springer.

Preprint. Final version is forthcoming in:
Philosophy of Science

Are higher mechanistic levels causally autonomous?

Peter Fazekas [1,2,✧] and Gergely Kertesz [3]

[1] *Centre for Philosophical Psychology, University of Antwerp, Belgium*
[2] *Cognitive Neuroscience Research Unit, CFIN, Aarhus University, Denmark*
[3] *Department of Philosophy, Durham University, UK*

✧corresponding author, email: fazekas.peter@gmail.com

Abstract

This paper provides a detailed analysis and explores the prospects of the arguments for higher-level causal autonomy available for the proponents of the mechanistic framework. Three different arguments (a context-based, an organisation-based, and a constraint-based) are distinguished. After clarifying previously raised worries with regard to the first two arguments, the paper focuses on the newest version of the third argument that has recently been revived by William Bechtel. By using Bechtel's own case study, it is shown that not even reference to constraints can establish the causal autonomy of higher mechanistic levels.

1. Introduction

The mechanistic approach aims at accounting for a target phenomenon — typically the behaviour of a higher-level whole — in terms of an underlying mechanism constituted by the organised activities of lower-level parts that jointly produce the very higher-level behaviour in question (Machamer et al. 2000; Bechtel and Abrahamsen 2005; Illari and Williamson 2012). Mechanistic explanations dominate the biological and life sciences (Craver 2007; Craver and Darden 2013), are also important in the physical and engineering sciences (Glennan and Illari, 2018), and are even claimed to be able to contribute to our understanding of causation (Glennan 1996, 2010, 2017). Nevertheless, the ontological commitments and metaphysical implications of the mechanistic framework are far from being clear (Fazekas and Kertész 2011; Soom 2012; Rosenberg 2015; Eronen 2015; Kaiser and Krickel 2016; Krickel 2019).

This paper focuses on the metaphysical implications of the mechanistic approach with regard to the *causal autonomy* of higher levels. Can higher-level wholes be autonomous with respect to the corresponding lower-level mechan-

isms underlying them and producing their characteristic behaviour? According to a wide consensus, they certainly can in an epistemological sense. But what about causal autonomy? Can higher-level wholes possess unique causal powers? This question is hotly debated. On the one hand, it has recently been argued that the causal autonomy of higher-level entities is incompatible with some core commitments of the mechanistic approach (Fazekas and Kertész 2011; Soom 2012; Rosenberg 2015). On the other hand, William Bechtel routinely presents the mechanistic framework as one that is well equipped to ensure the autonomy of higher-level entities even in a strict causal sense (Bechtel 2007, 2008, 2009, 2017a, 2017b; Bechtel and Abrahamsen 2008).

In this paper our aim is to advance this debate by providing a detailed analysis and exploring the prospects of the arguments for higher-level causal autonomy available for the proponents of the mechanistic framework. Mechanists argue for causal autonomy relying on an entangled mixture of different arguments, so the first goal of our endeavour is to disentangle the relevant parts of the literature, and to reconstruct the different arguments in play. We will distinguish three different arguments: a context-based, an organisation-based, and a constraint-based type that are independently motivated by different commitments.

The context-based and organisation-based arguments have played a central role at previous stages of the debate, and have been in the focus of recent objections (Fazekas and Kertész, 2011; Soom, 2012; Rosenberg, 2015). In his recent contributions, Bechtel tries to discredit these objections (Bechtel, 2017a, 2017b), so the second goal of our paper is to clarify how the context-based and organisation-based arguments can be answered.

In his most recent papers Bechtel also emphasises the importance of a third kind, a constraint-based argument for the causal autonomy of higher mechanistic levels (Bechtel, 2017a, 2017b). On the face of it, this line of thought is able to evade existing objections and bestow higher-level entities with unique causal powers. The third and major goal of our paper is to argue that this, however, is not the case. By using Bechtel's own case study, we will show that not even reference to constraints can establish the causal autonomy of higher mechanistic levels.

2. Background: mechanisms, levels and causal autonomy

The mechanistic programme targets a characteristic behaviour of an entity, and aims to describe the mechanism that produces the behaviour in question. Mechanistic explanations proceed via (1) identifying working parts (Bechtel 2008) or

components (Craver 2007) of the target entity, (2) describing the specific spatial and temporal organisation of the parts, and (3) demonstrating that the overall activity of the components organised in such a way is able to produce the behaviour in question (Machamer et al. 2000; Craver 2007; Bechtel 2008).

The spatially and temporally arranged operations of the parts constitute the behaviour of the target entity in the sense that the organised overall activity of the parts exhibit the target behaviour (Craver, 2007). That is, according to the logic of this framework, the mechanistic agenda is to find those components whose joint operations result in the very behaviour that the mechanistic programme aims at accounting for.

The mechanistic approach works with a multi-level picture: the target phenomenon is at a higher level while the entities that together produce the characteristic behaviour of the target are at a lower level (Craver 2007; Bechtel 2008). Levels are defined in terms of the working parts the organised activity of which constitutes the target phenomenon. Entities are at the same level if they are the working parts of the same mechanism: "[i]t is the set of working parts that are organised and whose operations are coordinated to realise the phenomenon of interest that constitute a level" (Bechtel 2008, 146). Similarly, entities are at different levels if one of them (the lower-level entity) is a working part of a mechanism that produces the behaviour characteristic of the other one (the higher-level entity): "X's ϕ-ing is at a lower mechanistic level than S's ψ-ing if and only if X's ϕ-ing is a component in the mechanism for S's ψ-ing" (Craver 2007, 189).

Since the behaviour of a component can similarly be analysed in terms of the organised activities of still lower level entities, the mechanistic framework presents the world as a nested hierarchy of mechanisms, in which entities residing at a lower level form a mechanism that produces the behaviour of a higher level entity that is a working part of a higher level mechanism, etc.

Those who argue for the causal autonomy of higher levels claim that higher levels are not just causally potent (which would be true even if the causal powers of higher-level entities were also the powers of certain lower-level entities; see e.g. the subset view: Shoemaker, 2007) but also possess unique causal powers. As it has lately been emphasised, such an argument for causal autonomy needs to start with establishing the claim that lower levels are causally not closed (Hendry 2010). William Bechtel (2017a, 2017b) has recently taken up the challenge and tried to demonstrate that the way higher level constraints and boundary conditions work disproves the causal closedness of lower levels, renders the mechanism as a whole not just causally potent, but also caus-

ally autonomous, and helps evade the causal exclusion argument (Kim 1998). Note that to support such a causal autonomy claim one needs to demonstrate that the causal influence a higher-level entity exerts to determine the unfolding of some lower-level events uniquely belongs to the higher-level, ie. it is not exerted by any lower-level entity.

3. Arguments for higher-level causal autonomy

Over the last two decades many attempts to argue for the causal autonomy of higher levels have been published, often presenting an entangled web of three different arguments.

3.1 The argument from context

An important consequence of how the mechanistic framework defines the criteria of being at the same level and being at different levels (see Sec. 2) is that levels of mechanisms are *local*: they are well-defined only within a given compositional hierarchy. If two entities are not working parts of the same mechanism, then there is no meaningful way to address the question whether they are at the same level (Craver 2007, 192; Bechtel 2008, 147). Due to this locality, lower levels are never causally closed, since lower levels are never extended beyond the set of entities that together constitute a mechanism, and thus they are always restricted and partial. So within the mechanistic framework there are no comprehensive lower levels that are causally complete and closed (Bechtel 2008, 148). Local levels restricted to the constituents of a mechanism typically lack resources to account for effects exerted from outside the mechanism, i.e. from the *context* in which the higher-level whole is embedded (Bechtel 2007, 183; 2008, 152). As a consequence, such lower level effects will have only higher level causes.

3.2 The argument from organisation

The spatial arrangement of the entities forming a mechanism and the temporal organisation of their activities are crucial determinants of a mechanism. However, just as resources necessary to account for contextual effects, resources required to account for arrangement and organisation are also unavailable at the lower level of the parts. The spatial and temporal structure of the parts is independent of their behaviour: parts conforming to the same laws and producing identical behavioral repertoire in isolation can nevertheless be organised into very different structures (Bechtel 2007, 183). This organisation is imposed upon

the parts by the higher level whole (Bechtel 2008, 150; 2009, 554-57), and thus is a manifestation of the unique causal powers of higher levels.

3.3 The argument from constraints

The spatial and temporal organisation of the parts constrains the behaviour of the parts: modes of organisation imposed on the constituents restrict how they can interact with each other (Bechtel 2009, 555-57). Similarly, the functioning of the mechanism as a whole in its higher-level context impinges specific conditions upon the mechanism that, at least partly, dictate how the parts can operate (Bechtel 2008, 240; 2009, 557-59). So by imposing a structure on the parts and by interacting with its environment the higher level whole constrains the behaviour of the parts — which is an extra causal influence on what happens inside the mechanism exerted by higher level entities, and thus a further source of the causal autonomy of higher levels (Bechtel 2017a, 271).

4. Clarifying the answers to the arguments from context and organisation

Recent criticisms of the idea that higher mechanistic levels are causally autonomous provided reasons to resist the arguments from context and organisation. It has been argued that the way the mechanistic framework thinks about constitution, causation and levels are incompatible with the main claims of these arguments.

4.1 Identifying higher and lower level causal roles

It is a fundamental tenet of the mechanistic framework that the organised activity of the constituent parts of a mechanism produces the very behaviour that characterises the target phenomenon, since this is the requirement that ensures the success of explaining a higher-level phenomenon mechanistically in terms of the behaviour and organisation of certain lower-level entities. If the behaviour produced by the organised activity of the lower-level entities was not identical to the characteristic behaviour of the target phenomenon, then the mechanistic story would clearly not be an account of the target phenomenon (Fazekas and Kertész 2011; Soom 2012; Rosenberg 2015). That is, the mechanistic framework is inherently committed to *identifying the causal roles* a whole plays at the higher-level with the causal roles the organised activity of the parts plays at the lower-level. So, even if causal processes at different levels look different, they are not different—in fact, the corresponding ones must be identical (Bechtel 2008, 2009; Fazekas and Kertész 2011).

Note that this picture is not eliminativist, for higher-level entities are token-identical with sets of lower-level ones. It is not epiphenomenalist either, for higher-level entities do possess causal powers — the same causal powers that are also possessed by certain spatially and temporally structured sets of lower level entities (see Fazekas and Kertész 2011; Soom 2012). This latter point needs special emphasis, given that Bechtel interprets his opponents as if they were arguing for the unjustified extreme view, which is "highly reductionist", "represents all activity at one lowest level" (Bechtel 2017a, 269) and renders "higher levels epiphenomenal" (Bechtel 2017a, 262; for such a view see Rosenberg 2015). Contrary to this, the picture advocated here is compatible with the usefulness of higher-level enquiries (as the only sources of higher level descriptions), and the importance and significance of higher-level accounts (as the only sources of certain generalisations, as say, in the case of multiple realisability); i.e. it is compatible with the epistemic or explanatory autonomy of higher levels. What it is incompatible with is the view Bechtel wants to argue for: the *causal* autonomy of higher levels.

4.2 Answering the argument from context

Another commitment explicitly endorsed by proponents of the mechanistic framework is that causation is an intra-level phenomenon: causal links do not span between different levels, they are confined to single levels. As Craver and Bechtel put it: "[t]here are no causal interactions beyond those at a level" (2007, 561; see also Craver 2007; Bechtel 2008, 2017; Fazekas and Kertész 2011). Note, however, that this commitment to causation as an exclusively intra-level phenomenon is in tension with the claim that levels are local. Due to locality, even if it is possible to analyse two entities that are interacting, causally connected parts of a mechanism in terms of lower-level sub-mechanisms, it is not possible to decide whether the lower-level parts constituting the sub-mechanism responsible for the behaviour of one of the entities are at the same level as the lower-level parts constituting the other sub-mechanism (Craver 2007).

The problem with this consequence of locality is the following: (i) in accordance with the logic of mechanistic explanations, the organised activity of the parts constituting the sub-mechanisms produce the very behaviours that are characteristic of the entities analysed; (ii) if an interaction between the two entities is part of this behaviour then there will be an interaction between the two sub-mechanisms as well; (iii) and if causality is strictly an intra-level phenomenon then the causal connection between the two sub-mechanisms entails that the

two sub-mechanisms are at the same level (Fazekas and Kertész 2011; Eronen 2015).

That is *lower levels can be extended* beyond the original scope of mechanistic enquiry. Consequently, lower levels are not necessarily restricted and incomplete, and thus there is no principled reason to think that they are causally not closed. In particular, note that the higher-level context that plays a crucial role in the 'argument from context' is the higher-level environment of the entities targeted by mechanistic decomposition. This environment consists of those higher-level entities that interact with the entity under scrutiny. Since lower levels can be extended exactly along the lower-level counterparts of such higher-level connections, contrary to the claim of the 'argument from context', resources are very much available at the lower-level to account for the effects exerted by the context of the mechanism (Fazekas and Kertész 2011, Soom 2012).

4.3 Answering the argument from organisation

Bechtel explicitly argues that how the components are spatially, temporally, and relationally organised go beyond the account of the parts and their operations. However, a lower-level account is not restricted to the (intrinsic) characteristics of the parts and the behavioural repertoire they produce in isolation. Thinking so would simply misidentify the supervenience base of the higher level entity under scrutiny. The organisation of the lower-level parts crucially determines what kind of behaviour the overall mechanism produces. So a lower-level account must include information about the spatial and temporal structure of the parts, and their interactions. In fact, the organisation of parts can only be accounted for at the lower-level, since it is the lower-level methodology that can target lower-level entities and thus can uncover their organisation, and it is the lower-level terminology that is apt for describing the parts and their structure. Note that these are criteria that Bechtel himself proposes as ways to characterise higher and lower levels: he claims that at higher and lower levels different experimental and explanatory strategies and different vocabularies are needed (Bechtel 2007, 185; 2008, 155-57). That is, since spatial, temporal and relational facts of lower-level entities are discovered by lower-level enquiries, and are characterised by lower-level vocabularies, by Bechtel's own standards they are very much integral parts of lower-level accounts (Fazekas and Kertész 2011).

5. Answering the argument from constraints

Bechtel (2017a) acknowledges that there is room for debate with regard to his approach to the autonomy of higher levels. As he admits, he and Craver have

"failed to bring out in what sense higher levels are involved in producing [...] effects at the lower-level" (Bechtel 2017a, 254). Bechtel aims to clarify this issue by emphasising the role constraints play in shaping lower-level causal processes, i.e. by providing a refined version of the 'argument from constraints'.

Bechtel claims that parts embedded in a mechanism behave differently than in isolation because of the restrictions imposed upon them (Bechtel 2017a, 271). These restrictions limit the behavioural repertoire of the parts by radically reducing the degrees of freedom available to them, thereby forcing them into a narrow behavioural space. Bechtel emphasises those cases in which restrictions are generated by feedback loops. Feedback makes the behaviour of the individual components sensitive to the actual state of the whole mechanism that they are embedded in. Since such a state can change in accordance with a characteristic dynamics, the observable behaviour of the parts can also change with time resulting in surprising patterns of different effects evoked by the same input (Bechtel 2017a, 272-273).

Note that on the face of it this strategy has the potential to avoid the worries raised with regard to the arguments from context and organisation. Even if the causal roles of higher-level wholes (the behaviours produced as responses to specific influences) are identical to the causal roles of sets of interacting lower-level parts constituting a mechanism, and even if lower levels are extendable along the causal connections of the lower-level entities constituting the mechanism to non-constituents outside the mechanism boundary, constraints can still restrict what can happen *inside* the mechanism. If such constraining effects can be attributed to no lower-level entities but only to the higher-level whole, then the 'argument from constraints' goes through, and the higher-level proves to be causally autonomous. (Here we follow Bechtel who argues that constraints are causal. For a different view, see, for example: Paoletti and Orilia 2017, 4-7.)

5.1 A case study: the circadian clock mechanism

Bechtel's own case-study to support the 'argument from constraints' is about the intra-cellular molecular 'clock mechanism' that is responsible for circadian rhythmicity — cyclic oscillations of daily behaviours and physiological functions. The molecular clock mechanism partly consists of a transcriptional-translational feedback loop involving so-called 'clock genes' *Per* and *Cry* and their protein products PER and CRY. Inside the nucleus, transcription factors BMAL1 and CLOCK drive the transcription of the genes *Per* and *Cry* to RNAs, which then are decoded by ribosomes to produce proteins PER and CRY. After a translocation back to the nucleus, PER and CRY inhibit the transcriptional

effects of BMAL1 and CLOCK through direct protein-protein interactions. Via this negative feedback loop PER and CRY autoregulate their own transcription, which results in a periodic increase and decrease in their concentration, and defines the phases of the oscillation of the clock mechanism. During the night PER and CRY concentration is high, while it is low during the day such that it peaks in the middle of the night after which it decreases and bottoms out in the middle of the day. Correspondingly, the level of *Per* and *Cry* transcription is low during the night, increases at dawn as PER and CRY concentration falls, at its maximum during the day and then decreases at dusk (the presentation here is simplified; for full details see Rosenwasser and Turek 2017). That is, the transcription of *Per* and *Cry*, which is a component of the mechanism, depends on the phase of the oscillation, and thus is sensitive to the actual state of the mechanism as a whole (Bechtel 2017a, 257). The constraining effects of the higher level whole on the behaviour of the components can most clearly be seen, Bechtel argues, in the different effects of light exposure on *Per* transcription at different times of the day: while light input has no effect during daytime, at dusk it delays, whereas at dawn it advances the phase of the oscillation (Bechtel 2017a, 267).

Bechtel's claim here is that characteristics of the lower-level mechanism, like the behaviour of certain parts (the level of transcription of *Per* and *Cry*), and how the organised activity of the parts processes a given input (exposure to light) are determined by the actual state of the higher-level whole (the phase of the oscillation). Recall that Bechtel uses this example to support his 'argument from constraints' against the causal closure of the lower-level and for the causal autonomy of the higher-level (see e.g. Bechtel 2017a, 272). So the claim that he really needs is that the way in which the higher-level whole determines the unfolding of lower-level events is via a causal influence, and that no lower-level entity exerts the same causal influence. Does Bechtel's own example support this claim?

5.2 Constraining effects are exerted by lower-level entities

Bechtel crucially relies on the phase of the oscillation of the circadian clock mechanism as the determinant of how lower-level processes unfold. The phase of the oscillation, however, is not a characteristic of some higher-level whole that has no counterpart at the lower level. On the contrary, the molecular description that characterises the clock mechanism can and does define the phases in purely lower-level (molecular) terms: as the periodic changes in the concentrations of PER and CRY (see above, also Rosenwasser and Turek 2017, 352-55).

So the phase of the oscillation is not some higher-level factor that causes changes in the level of transcription of *Per* and *Cry* in a way such that the relevant causal power is not possessed by any lower-level entity. It is the changes in the concentrations of PER and CRY that cause changes in the level of transcription of *Per* and *Cry* — via the direct protein-protein interactions with BMAL1 and CLOCK inhibiting their drive of the transcription of *Per* and *Cry*. Similarly, it is not the phase of the oscillation, as a higher-level entity per se, that constrains how the parts of the mechanism change their behaviour as a response to a light signal, but the actual level of concentrations of PER and CRY (see Bechtel 2017a, 267-68).

Interestingly, Bechtel is in agreement with us with regard to the presence and importance of these lower-level interactions. He acknowledges that "the phase of the oscillator at a time *just is* the concentrations of PER, CRY" (Bechtel 2017a, 257, our emphasis). And he also acknowledges that "[a]s a result of the interconnectivity of the parts, especially the feedback loops, the [...] mechanism functions as a unit, with the operations of the individual parts of the mechanism *determined by other parts of the mechanism*" (Bechtel 2017a, 268, our emphasis). Nevertheless, he thinks that these interactions are compatible with the causal autonomy of higher levels (Bechtel 2017a, 272). But they are not. These interactions, by ensuring that the lower-level effects in question have lower-level causes, causally close the lower-level with regard to the behaviours under scrutiny, and leave no room for the causal autonomy of the higher-level.

Per, *Cry*, PER, CRY, BMAL1, CLOCK, etc. are interacting working parts of the clock mechanism, and as such — by the standards of Bechtel's own definition that puts working parts of a mechanism at a lower level than the whole the behaviour of which is produced by the mechanism (Bechtel 2008; see also Craver 2007) — they are lower-level entities. Furthermore, they, their interactions, their spatial arrangement and the temporal organisation of their activities are all studied by lower-level methodologies, and characterised by lower-level vocabularies (see Sec. 4.3). Is there contextual information that cannot be captured at this level? No: the relevant contextual information concerns the effect of light on the phase of the clock mechanisms — but this is also defined in terms of the lower level, as the boosting effect of light on *Per* and *Cry* transcription. What is left? Nothing. All information is there at the lower level, and all causal influences can be exerted by lower-level entities. The higher level possesses no unique causal powers that couldn't be attributed to some lower-level factor. Therefore, the higher-level is not causally autonomous.

Bechtel is right that feedback loops are important. But they are important as parts of lower-level interactions. Having an account of how feedback loops actually work reveals how lower-level processes impose complex constraints upon their own unfolding, and thus, instead of disproving, it contributes to appreciating the causal closedness of the lower level.

6. Conclusion: higher mechanistic levels are not causally autonomous

Here we distinguished three different arguments for the causal autonomy of higher mechanistic levels: the arguments from context, organisation and constraints. Upon closer reflection, it is evident that the constraining effects that restrict and determine the behaviours of lower-level parts are brought about by the interactions between lower-level entities. Parts embedded in a mechanism behave differently than in isolation because they are in constant interactions with other parts of the mechanism and also with further entities external to the original mechanism. These interactions internal and external to a given mechanism are what define the organisation of the parts and the context of the mechanism, respectively, thereby encoding both organisation and context based constraints at the lower-level. The effects of context, organisation and constraints can all be accounted for in terms of the causal influences of lower level entities and activities. That is, within the mechanistic framework, the causal autonomy of higher levels cannot be established.

References

Bechtel, William. 2007. “Reducing psychology while maintaining its autonomy via mechanistic explanation." In *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience and Reduction*, ed. Maurice Schouten and Huib Looren de Jong, 172-198. Oxford: Basil Blackwell.

———. 2008. *Mental mechanisms: philosophical perspectives on cognitive neuroscience*. New York: Psychology Press.

———. 2009. “Looking down, around, and up: Mechanistic explanation in psychology.” *Philosophical Psychology*, 22:543-564.

———. 2017a: “Explicating top-down causation using networks and dynamics.” *Philosophy of Science*, 84:253-274.

———. 2017b: “Top-down causation in biology and neuroscience: Control hierarchies.” In *Philosophical and scientific perspectives on downward causation*, ed. Michele Paolini Paoletti and Francesco Orilia. London: Routledge.

Bechtel, William, and Adele Abrahamsen. 2008. “From reduction back to higher levels." In *Proceedings of the 30th Annual Conference of the Cognitive Science Soci-*

ety, ed. B. C. Love, K. McRae and V. M. Sloutsky, 559-564. Cognitive Science Society.

Craver, Carl. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Craver, Carl, and William Bechtel. 2007. "Top-down causation without top-down causes." *Biology and Philosophy*, 22:547-563.

Craver, Carl, and Lindley Darden. 2013: *In Search of Mechanisms. Discoveries across the Life Sciences*. Chicago: University of Chicago Press.

Eronen, Markus. 2015. "Levels of organization: a deflationary account." *Biology and Philosophy*, 30:39-58.

Fazekas, Peter, and Gergely Kertész. 2011: "Causation at different levels: tracking the commitments of mechanistic explanation." *Biology & Philosophy*, 26:365-383.

Glennan, Stuart. 1996. "Mechanisms and the nature of causation." *Erkenntnis*, 44:49–71.

———. 2010. "Mechanisms, causes, and the layered model of the world." *Philosophy and Phenomenological Research*, 81:362-381.

———. 2017. *The new mechanical philosophy*. Oxford: Oxford University Press.

Glennan, Stuart, and Phyllis Illari. 2018. *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. New York: Routledge.

Illari, Phyllis, and Jon Williamson. 2012. "What is a mechanism? Thinking about mechanisms across the sciences." *European Journal for Philosophy of Science*, 2:119-135.

Kaiser, Marie I., and Beate Krickel. 2016: "The metaphysics of constitutive mechanistic phenomena." *The British Journal for the Philosophy of Science*, online first, 1–35.

Kim, Jaegwon. 1998. *Mind in a physical world*. Cambridge, MA: MIT Press.

Krickel, Beate. 2019. *The Mechanical World – The Metaphysical Commitments of the New Mechanistic Approach.* Springer.

Machamer, Peter, Lindley Darden, and Carl Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science*, 67:1–25.

Paoletti, Michele Paolini and Francesco Orilia. 2017. *Philosophical and Scientific Perspectives on Downward Causation*. New York, NY: Routledge.

Rosenberg, Alex. 2015. "Making mechanism interesting." *Synthese*, Electronically published May 30. doi:10.1007/s11229-015-0713-5.

Rosenwasser, Alan M. and Fred W. Turek. 2017. "Physiology of the Mammalian Circadian System". In *Principles and Practice of Sleep Medicine*, ed. Meir. H. Kryger, Thomas Roth, and William C. Dement, 351-361. Elsevier Health Sciences.

Shoemaker, Sydney. 2007. *Physical Realization*. Oxford: Oxford University Press.

Soom, Patrice. 2012. "Mechanisms, determination and the metaphysics of neuroscience." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43:655-664.

Reframing the Homology Problem
Devin Gouvêa

Abstract

Recent philosophical work on biological homology has generally treated its conceptual fragmentation as a problem to be solved by new accounts that either unify disparate approaches to homology or specify sharp constraints on its meaning. I show that several proposed solutions either misunderstand or ignore central features of comparative biological research, despite attempts to capture scientific practice. I conclude that the problem is incorrectly framed and that disagreements about homology may be epistemically fruitful. Empirically tractable debates are more likely to occur among biologists who share theoretical perspectives on homology. Philosophers should consider homology not merely as a generator of inductive generalizations but also as a scaffold for meaningful empirical comparisons.

1. Introduction

"I will grant that someone might be able to generate an original thought concerning homology, but I doubt it." So complained the herpetologist David Wake nearly twenty years ago, during a revival of biological interest in the topic. Wake certainly did not doubt the importance of homology—a slippery notion perhaps most neutrally defined as correspondence between the parts of different organisms (Brigandt 2012). On the contrary, Wake elsewhere proclaimed it to be "the central concept for all of biology" (1994, 268). Having established this bedrock position for homology, however, Wake thought that continued discussion of its meaning was a distraction from more interesting biological research questions. "Isn't it time to move on?" he asked (1999, 24).

Wake's caution notwithstanding, speculation about the meaning of homology has continued apace in both biological and philosophical circles. Philosophical attention to the topic has been influenced by the rising tide of interest in scientific practice. In the first few sections of this paper, I will briefly review and critique several philosophical analyses of homology that appeal to some aspect of scientists' aims or methods. They exhibit two general approaches — some offer restrictive accounts of homology that deliberately exclude certain biological positions, while others offer compatibilist accounts that reconcile these positions. While the latter are more successful, both kinds of approach ultimately fail to capture important aspects of biological practice.

Given the diversity of biological practice, the pervasiveness of homology, and the broad theoretical level at which different accounts are traditionally characterized, this failure is not surprising. In response, I suggest that philosophers need to reconsider

whether conflict between theoretical accounts of homology is really such a problem after all. Such conflict can coexist with broad agreement on the underlying methodological principles that support the reconstruction of evolutionary history. In contrast, biologists working within the same theoretical perspectives often pursue extended conflict about empirically tractable questions for which the data is still too limited, or interpretations still too underdetermined, to settle the matter. Homology is therefore just as much a tool for generating provocative comparisons as it is for supporting inductive generalizations based on natural kinds.

2. Homologizing as Kinding

Catherine Kendig (2016) offers a restrictive account of homology that is particularly emphatic about attending to practice. Her goal is to shift the focus of the debate away from "defining *homology*" to "the practices of *homologizing*" (106). She takes homology to be a natural kind concept, and *homologizing* to be a set of rule-following practices, or "kinding activities that have shaped, and continue to shape, the meaning and use of *homology*" (106–7). The first part of her paper analyzes the long history of comparative practices, from the comparative anatomical investigations of Vesalius and Belon, through Richard Owen's attempts to reconcile Cuvier's emphasis on functional unity with Geoffroy's universal body plan, to Darwin's reinterpretation of abstract archetypes as causally efficacious ancestors. The message of this history is that "[t]he concepts used within comparative biology and the activities of natural kinding have a history of being revised and retuned in response to comparative research practices" (118).

Thus far, Kendig seems poised to champion a pluralist account of homology concepts. In her account of the twentieth century, however, she switches gears to champion a particular notion of homology. She opposes the cladistic practice of mapping homology onto monophyletic groups without acknowledging its particular aim, namely to provide reliable classifications and historical hypotheses (113). "Homologizing as monophyleticizing" is an "all-or-nothing" approach that ignores all traits which are not inherited through a continuous ancestral lineage and "vociferously" objects to partial homology as a "threat to the Modern Synthesis" (115).[1] Against this foil, Kendig claims that "practices of kinding in comparative biology are reshaping the conception(s) of homology" (117). These practices, drawn from developmental and organismal biology, reveal phenotypic traits to be "mosaic" composites of modular units that can be rearranged in a combinatorial fashion during evolution. Kendig also takes symbiosis to be a source of variation that transcends individual genetic inheritance. She concludes by arguing that "multidimensional homology thinking" has replaced standard evolutionary accounts, presumably in much the same way that "[t]he historical notion of Darwin's 'ancient progenitor' replaces Owen's idealist 'archetype'" (113).

While Kendig highlights some underappreciated features of phenotype change over time, her privileging of multidimensional homology thinking over cladistic

[1] The quote that supports this point—Donoghue's claim that "partial homology is incompatible with standard evolutionary views" (1992, 172)—seems to be taken out of context. He is referring to standard views of homology, not evolutionary theory in general.

approaches is not sufficiently motivated, and she fails to consider alternative traditions that may be more congenial to her view. For example, Wake (1999) readily admits the existence of partial homology, and Brian Hall (2003) proposes a continuum between homology and homoplasy defined by the differential conservation of developmental resources and phenotypic traits (see section 4 below for more on his view).

3. Must Homology and Homoplasy Be Kept Apart?

Adrian Currie (2014) also uses scientific practice to motivate a restriction on homology concepts, but with opposite results to Kendig. Whereas she rejects the cladists' sharp separation of homology from homoplasy (roughly, biological similarity without whatever kind of correspondence is considered necessary for homology), Currie embraces this distinction as necessary to make sense of practice. His methodology also appears more promising. While Kendig is selective in her assessment of contemporary homologizing, Currie claims to have identified four epistemic roles that are ubiquitous in biology and for which a sharp distinction between homology and homoplasy is essential.

Across these diverse situations, argues Currie, biologists use the distinction between homology and homoplasy when distinguishing signal from noise, or "splitting evidential wheat from chaff" (704). Which one is which may depend on the situation, but the need for a strict distinction remains. For Currie, the distinction must have a genealogical foundation[2]; "two similar traits [in different lineages] are homologous just

[2] Currie uses the term "taxic" interchangeably with "genealogical" and "phylogenetic." See section 6 for an argument that this is misleading.

in case they are present in the most recent common ancestor; homoplastic just in case they are *not* present in the common ancestor." By contrast, a developmental approach to homology would identify traits as homologous "just in case they are the products of the same developmental process" (702).[3] Currie allows that there may be some traits which are neither homologous nor homoplastic, but there must not be any overlap—no trait can be both homologous and homoplastic.

In this section, I will briefly review these roles and the particular biological case that illustrates them. I accept that Currie has identified an epistemically important distinction, but dispute that it concerns homology and homoplasy. To reinforce the point, the following section looks more closely at how the relationship between development and genealogy is construed by Brian Hall, the main foil for Currie's account, and Günter Wagner, champion of the most worked-out developmental theory of homology.

First, Currie claims that the distinction is essential for determining phylogenetic relations in the first place. Similarity of morphological or molecular features is essential to infer these relationships but biologists have long recognized that not all similarities are equally informative. For example, distinct but related lineages may retain enough common developmental and genetic heritage that they respond to selection in similar ways. Systematists disagree as to whether these misleading characters (identified as homoplastic) can sometimes be recognized in advance of cladistic analysis (and thus excluded from consideration) or whether they can only be revealed by the topology of a

[3] Proponents of such accounts might object to this definition, which overstates the developmental similarity required for homology.

completed tree. Currie mentions both possibilities without clearly distinguishing them. He first identifies "diagnoses of homology and homoplasy" as the *result* of "statistical analysis of patterns of similarity" (710) and later describes them as different kinds of *input* — "homologies count as data-points for common ancestry, while homoplasy is noise" (711). Both are obviously phylogenetic applications of the distinction between homology and homoplasy, and in either case Currie could argue that allowing overlap between these categories would confound their epistemic roles.

Second, according to Currie, biologists need the distinction when they use analogical reasoning to infer the traits of inaccessible organisms from those that are better characterized. Extinction is one cause of inaccessibility; others include extreme habitats, practical constraints, and ethical concerns. In all these cases, some features can still be known but others remain beyond reach. Unobserved traits are often attributed to the (inaccessible) *target lineage* by appealing to a (better known) *model lineage* that exhibits the trait of interest alongside other some other characteristic(s) known to be shared by both lineages. What justifies the projection from the coupling of traits in one lineage to their coupling in another lineage? According to Currie, homologous and homoplastic relationships answer this question differently and thus must be kept apart.

If the trait were present and coupled in the common ancestor of the model and target, then the inference is justified by an appeal to the stability of inheritance. If the trait was not present in the common ancestor, we have to appeal to a different kind of regularity, one grounded in the similarity of selective regimes. In both cases, there are additional factors to consider. We may be cautious about inferring the continuous

inheritance of traits that are especially labile or between lineages that are especially distant. Likewise, strong selection in related lineages may increase the probability of parallel evolution. Not all cases of homology and homoplasy can ground the inference, but both provide important evidence for such an inference. The evidence, however, is of fundamentally different types, and so we should keep these two concepts distinct.

A third reason to maintain the distinction is that it aids in testing adaptive hypotheses. The independent appearance of some trait in two lineages — an example of parallel evolution, which Currie classifies as homoplasy — can furnish evidence of adaptive function, particular when the environments are similar. But if the trait was not independently acquired, any adaptive hypothesis must first consider the ancestral environment and the original function of the trait.

Finally, Currie argues that evolutionary developmental biology, with its interest in evolutionary novelties, needs at least a derivative form of the delineated genealogical account. Under one definition, a novel trait is just one that has no homologue in any ancestral taxon. This is certainly a phylogenetic definition, but it does not require any particular contrast with homoplasy. The concept of novelty is itself rather vexed (Brigandt 2012) so this example provides perhaps the weakest support for Currie's claim.

Currie illustrates the example by referring to the dispute over a remarkable hypothesis that the birdlike dinosaur *Sinornithosaurus* was venomous. Gong et al. (2010, 2011) advance this hypothesis on the basis of particular morphological traits and analogies with extant venomous taxa. Gianechini et al. (2011) dispute their interpretation of both the anatomical and the phylogenetic evidence. According to Currie,

understanding this exchange requires "requires contrasting homoplasy and homology along taxic lines—to make sense of the dispute we need the distinction" (707). Without going into details of the dispute, I will present four questions that exemplify Currie's four epistemic roles for the distinction between homology and homoplasy.

1) What is the relationship between the theropod clade to which *Sinornithosaurus* belongs and other major dinosaur clades? 2) Was *Sinornithosaurus* venomous? 3) Are *Sinornithosaurus* fangs an adaptation to deliver venom to feathered prey or were they selected for some other function (or not directly selected at all)? 4) At what point did venom first evolve in the lineage leading to *Sinornithosaurus* — in other words, when and how did the evolutionary novelty arise?

In order to answer all four questions, biologists must know something about how venom and its anatomical correlates are distributed on the phylogenetic tree leading to *Sinornithosaurus.* 1) In reconstructing *Sinornithosaurus* ancestry, some morphological traits will be better indicators than others. Likewise, a solid tree will constrain our hypotheses about the evolution of traits like venom. 2) The analogy between the coupling of morphological traits in venomous lizards and snakes and their alleged coupling in *Sinornithosaurus* will be justified differently depending on whether or not the traits were present in a common ancestor. 3) If venom was present in the common ancestor of *Sinornithosaurus* and extant venomous taxa, we need to consider its adaptation to the ancestral environment. If it is a parallel evolution, we can more confidently analogize the ecological functions of venom. 4) Identifying the evolutionary novelty depends on which precursor traits the ancestor possessed.

Currie thus clearly illustrates that different phylogenetic patterns allow different kinds of inference, but it is not necessary to cash out these distinctions in terms of the contrast between homology and homoplasy. To reinforce this point, I turn to proponents of a developmental account.

4. Developmental Perspectives on Phylogeny

Brian Hall, one of the founding figures of evolutionary developmental biology, has insisted that homology and homoplasy should be understood as elements of a continuum. Does this view stand in tension with Currie's emphasis on phylogenetic clarification of ancestral relations? I argue that it does not. While Hall does indeed give developmental mechanisms a role in assessing homology, he remains adamant that they are insufficient for this purpose. In fact, questions about their significance for homology "are best posed—perhaps can only be posed—within the context of a sound phylogenetic analysis. Questions of mechanisms are second to phylogeny when assessing homology or homoplasy" (2007b, 476). The secondary place of development mechanisms reflects their complicated relationship with phenotypic evolution. Development can diverge even as a phenotypic trait is continuously inherited, and the phenotypic output of a conserved developmental mechanism can change over time.

"The history of life has been descent with modification" (Hall 2003, 427). For Hall, this unitary process underwrites a continuum between homology and a collection of relationships traditionally grouped under the heading of homoplasy.

Whether we are examining homoplasy (convergence), parallelism, reversals,

> rudiments, vestiges, atavisms or homology, we are dealing with common descent with varying degrees of modification of features as a result of natural selection tinkering with the genetic and developmental bases responsible for producing those features (ibid).

This passage reflects Hall's argument that only convergence—the evolution of similar traits in independent lineages—should be understood as truly homoplastic. However, he recognizes that independence cannot be precisely defined since *all* taxa share an evolutionary history that in many cases leads to conservation of genetic and developmental processes across great phylogenetic distances (2007a, 437–8).

The main difference between Hall and Currie, then, is simply that Hall recognizes parallel evolution as a type of homology because it depends on shared developmental resources. He still distinguishes this category from traditional homology, in which the trait itself is conserved along the ancestral lineage. Why does the distinction matter, in his view? Without emphasizing the affinity between homology and the other phenomena, he worries, we will be inclined to "search for different developmental and genetical mechanisms" and thus neglect the implications of shared evolutionary history. (2007b, 442). Rather than neglecting the importance of common ancestry, Hall places it at the base of his developmental account.

Hall's respect for genealogical approaches to homology led the systematist Joel Cracraft (2005) to count him as a "phylogenist" in his critique of evo-devo approaches to homology. The case of Günter Wagner, the originator and current champion of a developmental approach, therefore provides an instructive contrast. Wagner's original

articulation of the "biological homology concept" (BHC) made no mention of phylogeny and emphasized only shared developmental constraints (MacLeod 2011). But Wagner (1999) explicitly recognizes the importance of phylogeny in constraining mechanistic investigation. The initial steps in his early proposal for testing the BHC depend on phylogenetic analysis — putative homologues should be identified within two different but related taxonomic groups, and their distribution mapped onto a phylogenetic tree, ideally one constructed independently with molecular data. His recent (2014) book-length development of this approach is replete with phylogenetic trees and full of references to phylogenetic distributions that constraint the set of mechanistic hypothesis for the individuation and evolution of characters.

This sketch of Hall and Wagner gives us no reason to doubt that they would accept each of Currie's epistemic functions. They could maintain their differing views of homology by arguing that these examples, while sometimes framed in terms of homology, do not exhaust the meaning of the term.

5. Compatibilist Solutions

Given the failure of these two attempts to mount a practice-based restriction on homology, we might consider other approaches that emphasize the compatibility of different concepts.

Griffiths (2007) rejects the assumption "that principles of classification that can unify diverse particulars into broad categories…must be derived from our best explanatory theories of the domain to be classified" (655). Roughly, competing

"definitions" of homology are best understood as complementary explanations for a broad set of homology *phenomena* that are recognizable apart from those definitions. Operational criteria for recognizing homology, particularly the relative position of the parts and the existence of intermediates between them, have remained relatively constant since the nineteenth century. Both the genealogical and developmental approaches offer causal explanations of the phenomena of homology, the former in terms of common descent and the latter in terms of shared mechanisms.

Brigandt (2009) also emphasizes the compatibility of the two approaches, which "simply address different aspects and temporal stages of one complex phenomenon" (89). The unity of this complex phenomenon is provided by the HPC (homeostatic property cluster) view of natural kinds. Assertion of a homology relation between body parts in different lineages picks them out as members of a kind united by the homeostatic mechanisms that determine their individuality as units of phenotypic evolution. These mechanisms are in turn genealogically related in patterns that are traced by the methods of phylogenetic reconstruction. The developmental approach to homology emphasizes the individuating mechanisms, and the genealogical approach emphasizes their evolutionary relationships.

This view depends on a particular theoretical concept of biological characters as modular, quasi-independent units individuated by developmental genetic control mechanisms. There is good evidence (Wagner 2014) that such mechanisms exist for many body parts, and that they can change their component parts while maintaining their individuating potential. In such cases the two accounts may indeed be related by the

natural kind view. But what happens when the morphological characters proposed as homologous are not individuated in this way? Or when the lineages are extinct and the mechanisms are inaccessible? In these situations the unity seems likely to break down.

I am convinced by the argument of MacLeod (2011, 2013) that the natural kind picture glosses over importance methodological differences between the two approaches. Those characters which are most informative for reconstructing phylogenies will not be the most informative for understanding the developmental individuation of parts. A systematist, for example, will seek synapomorphic characters that uniquely diagnose all the descendants of a common ancestor. A developmental biologist, on the other hand, may be more interested in underlying mechanisms that are shared across groups, and thus homoplastic by cladistic reckoning. The account of Griffiths (2007) fares better on this analysis since it allows biologists to have different explanatory aims, but it does not account for the fact that proponents of different theoretical accounts see themselves as *identifying* homology, not merely explaining it. The two phases cannot be separated as neatly as Griffiths supposes.

6. Which Conflicts Should We Capture?

So far I have argued that a set of philosophical responses to conflicts over the meaning of homology—differing in their approaches but united in their concern to represent practice—fail to capture practice in important ways. The more restrictive accounts either simply ignore important epistemic functions (Kendig) or incorrectly assume that other important functions constrain empirical and theoretical research more than is actually the

case (Currie). The more compatibilist accounts do a better job of accounting for the range of approaches to homology, but they still gloss over important methodological differences both within and between the two main branches.

Given the immense diversity of scientific practices, philosophical accounts must face the challenge of determining how to individuate those practices. With sufficiently careful scrutiny, we might find nearly as many approaches to homology as there are individual scientists. Which ones should we try to capture? Recent philosophical work uniformly identifies a broad dichotomy between approaches focused on history and approaches focused on development. Is this the best way to frame the problem? Like earlier dichotomies—reviewed by Roth (1994, 303)—this one maps roughly onto the disciplinary divisions between systematists and biologists of other disciplinary persuasions. But a closer examination of those earlier dichotomies shows some uncertainty in how to count the categories. The original version of taxic homology, for example, deliberately broke with the requirement that homologous parts be traceable through "transformation series" to parts in common ancestors (Patterson 1982, Donoghue 1992).

I suggest we avoid such difficulties entirely by focusing our attention elsewhere, on the research practices that undergird the different theoretical accounts of homology (however we count them). My analysis of Currie (2014) shows that biologists with radically different theoretical accounts of homology might nevertheless agree that phylogenetic patterns constrain inferences about ancestral traits and adaptations in particular ways. They could likewise agree that not all traits are equally informative for

constructing evolutionary trees in the first place. Insofar as they disagree about what makes those traits novel, those disagreements do not turn their differing views about homology.

I suggest that we are more likely to find philosophically interesting disagreements among biologists who *share* broad theoretical approaches to homology. The paleontologists arguing over *Sinornithosaurus* venom do not reveal their hand on this topic (in fact they never explicitly discuss either homology or homoplasy) but given their subject matter, it seems a safe assumption that they will not consider developmental genetic individuation of body parts to be necessary for identifying homology. On the other hand, they disagree mightily about the interpretation of many empirical details. Are the teeth of *Sinornithosaurus* really as elongated as they appear, or just displaced from their sockets? Is one particular cavity in its skull, allegedly specialized to hold a venom gland, really anatomically separate from a neighboring cavity? Does the recent discovery of venom in new lizard and mammal taxa raise the plausibility of finding venom in ancestral dinosaurs?

To take another example from across the disciplinary aisle, evolutionary developmental biologists are currently engaged in a lively debate as to how the five digits of ancestral tetrapods gave rise to the three digits found in bird wings. The idea that developmental mechanisms delineate evolutionary units is implicit in this debate, so its participants are united in taking a broadly developmental approach to homology. The experimental evidence has ruled out certain simple scenarios in which the pattern of developmental control remained constant even as some digits were lost. However, it has

so far been insufficient to determine which of several plausible complex transformations have actually occurred (Young et al. 2011, Larsson and Wagner 2012). Some have therefore doubted whether digits are really developmentally individuated in the first place (Wagner 2014).

For Wagner, this kind of empirical ferment is a sign of the success of his view of homology. "The developmental account makes stronger assumptions about biological reality" than other approaches which do not postulate specific mechanisms for the maintenance of part individuality "and, thus, leads to testable predictions" (2014, 75). In general, he takes a pragmatic notion of homology concepts that prioritizes empirical fruitfulness over precise definitions. "*Any concept is only as good as the research program it inspires*" (245).

This sentiment is reflected by biologists from other perspectives. Though Donoghue (1992) worried that early versions of Wagner's program were overly narrow, he pointed out that "[a]chieving consistency with every version of homology may yield a definition that is of little use to anyone (179). De Pinna (1991, 368) argues that "an evaluation of definitions of homology acquires sense only against a specific frame of reference" defined by "a more encompassing method or theory," so that definitional disputes only have meaning against the backdrop of certain common assumptions. David Wake puts the point most strongly, in a paper arguing that the homology debate is a "distraction" from real research questions.

> I want to get on with it and to leave behind debates that started when biologists really did not have sufficient biological knowledge to appreciate the causes of biological similarity and when they did not yet understand that Darwin was right

in his view that there is one genealogy for all of life (1999, 45).

7. Conclusions

What would it mean for philosophers to follow Wake's lead and get on with it—to move past attempts to adjudicate or supersede debates about the meaning of homology? Various compatibilist approaches have taken a step towards this aim by characterizing the ways in which different theoretical accounts of homology are useful for different purposes (e.g. Brigandt 2003, 2012). We still need much more careful attention to the types of arguments that homology concepts, in their various forms, make possible, and we need to move away from the assumption that conflict between approaches is something to be explained away. This conflict is a natural consequence of the complexity of the phenomena under study and mirrors the disciplinary specialization necessary for propagating empirically successful techniques.

I am not saying, however, that we should simply defer to scientists when giving account of homology. Though metaphysical or definitional unification has brought only limited success, philosophical work has only scratched the possibilities for identifying some kind of epistemic commonality among the various homology practices. Several authors (Griffiths 2007; MacLeod 2011, 2013) have emphasized the role that homology plays in creating meaningful categories that can be subsequently used for inductive generalizations (for example from mouse to human physiology). This is one important function, but we should also consider the ways in which homology facilitates contrastive reasoning—the identification of meaningful differences between comparable individuals

(organisms, body parts, gene networks, etc.)—and the underlying causes of those differences. This would be a worthy application of the growing enthusiasm for practice-centered philosophy of science.

References

Brigandt, Ingo. 2003. “Homology in Comparative, Molecular, and Evolutionary Developmental Biology: The Radiation of a Concept.” *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 299: 9–17.

———. 2009. “Natural Kinds in Evolution and Systematics: Metaphysical and Epistemological Considerations.” *Acta Biotheoretica* 57: 77–97.

———. 2012. “The Dynamics of Scientific Concepts: The Relevance of Epistemic Aims and Values.” In *Scientific Concepts and Investigative Practice*, ed. Uljana Feest and Friedrich Steinle, 75–103. Berlin: De Gruyter.

Cracraft, Joel. 2005. “Phylogeny and Evo-Devo: Characters, Homology, and the Historical Analysis of the Evolution of Development.” *Zoology* 108: 345–56.

Currie, Adrian Mitchell. 2014. “Venomous Dinosaurs and Rear-Fanged Snakes: Homology and Homoplasy Characterized.” *Erkenntnis* 79: 701–27.

———. 2016. “Ethnographic Analogy, the Comparative Method, and Archaeological Special Pleading.” *Studies in History and Philosophy of Science* 55: 84–94.

Donoghue, Michael J. (1992). “Homology,” In *Keywords in Evolutionary Biology*, ed. Evelyn Fox Keller and Elisabeth A. Lloyd, 170–79. Cambridge: Harvard University Press.

Ereshefsky, Marc. 2009. “Homology: Integrating Phylogeny and Development.” *Biological Theory* 4: 225–29.

Gianechini, Federico A., Federico L. Agnolín, and Martín D. Ezcurra. 2011. “A

Reassessment of the Purported Venom Delivery System of the Bird-Like Raptor *Sinornithosaurus*." *Paläontologische Zeitschrift* 85: 103–7.

Gong, Enpu, Larry D. Martin, David A. Burnham, and Amanda R. Falk. 2010. "The Birdlike Raptor *Sinornithosaurus* Was Venomous." *Proceedings of the National Academy of Sciences of the United States of America* 107: 766–68.

———. 2011. "Evidence for a Venomous Sinornithosaurus." *Paläontologische Zeitschrift* 85: 109–11.

Griffiths, Paul E. 2007. "The Phenomena of Homology." *Biology & Philosophy* 22: 643–58.

Hall, Brian K. 2003. "Descent With Modification: The Unity Underlying Homology and Homoplasy as Seen Through an Analysis of Development and Evolution." *Biological Review* 78: 409–33.

———. 2007a. "Homology and Homoplasy." In *Handbook of the Philosophy of Science: Philosophy of Biology*, ed. Mohan Matthen and Christopher Stevens, 429–53. Amsterdam: North Holland.

———. 2007b. "Homoplasy and Homology: Dichotomy or Continuum." *Journal of Human Evolution* 52: 473–79.

Kendig, Catherine. (2016). "Homologizing as Kinding," In *Natural Kinds and Classification in Scientific Practice*, ed. Catherine Kendig, 106–28. New York: Routledge.

Larsson, Hans C. E. and Günther P. Wagner. 2012. "Testing Inferences in Developmental Evolution: The Forensic Evidence Principle." *Journal of Experimental*

Zoology Part B: Molecular and Developmental Evolution 318: 489–500.

MacLeod, Miles. 2011. "How to Compare Homology Concepts: Class Reasoning About Evolution and Morphology in Phylogenetics and Developmental Biology." *Biol Theory* 6: 141–53.

———. 2013. "Limitations of Natural Kind Talk in the Life Sciences: Homology and Other Cases." *Biological Theory* 7: 109–20.

Patterson, Colin. 1982. "Morphological Characters and Homology." In *Problems of Phylogenetic Reconstruction*, Systematics Association Special Volume 25, ed. K. A. Joysey and A. E. Friday, 21–74. London: Academic Press.

Roth, V. Louise. 1994. "Within and Between Organisms: Replicators, Lineages, and Homologues." In *Homology: The Hierarchical Basis of Comparative Biology*, ed. Brian K. Hall, 302–33. San Diego: Academic Press.

Wagner, Günter P. (1999). "A Research Programme for Testing the Biological Homology Concept," In *Homology*, ed. Gregory R. Bock and Gail Cardew, 125–40. Chichester, UK: John Wiley & Sons, Ltd.

——— (2014). *Homology, Genes, and Evolutionary Innovation*. Princeton: Princeton University Press.

Young, Rebecca L., Gabe S. Bever, Zhe Wang, and Günter P. Wagner. 2011. "Identity of the Avian Wing Digits: Problems Resolved and Unsolved." *Developmental Dynamics* 240: 1042–53.

Methodology at the Intersection between Intervention and Representation

Vadim Keyser[1]

Abstract: I show that in complex methodological contexts, representational and intervention-based roles require re-conceptualization. I analyze the relations between representation and intervention by focusing on the role of intervention in *mediating* representations. To do this, first I show how applied scientific practice challenges the simple distinction between representational and intervention-based roles of experiment/measurement. Then I discuss the complex interaction between representation and intervention applied to methodology in biomarker measurement.

1. Introduction

The relationship between intervention and representation is currently resurfacing in philosophy of science. Analytical treatments of the specific intersections between *representation* and *intervention* have recently been explored in Hacking (1983), Radder (2003), Heidelberger (2003), van Fraassen (2008), and Keyser (2017). These accounts analyze intervention-based experimental and measurement practice and the *consequences* for representing and model-building. Of particular interest in my discussion is that some of these accounts explicitly differentiate between representational and productive roles in scientific practice. For example, Heidelberger (2003) and van Fraassen (2008) discuss the representational and productive roles of instruments in experiment and measurement. In the former role, relations in a natural phenomenon are represented in an instrument (van

[1] California State University, Fresno. Email: vkeyser@csufresno.edu

Fraassen 2008, 94). In the latter role, instruments create new phenomena or mimetic phenomena, which resemble natural phenomena. Keyser (2017) takes the distinction between representation and production a step further to differentiate two types of experimental/measurement methodologies:

> When scientists measure/experiment they can *take* measurements, in which case the primary aim is to represent natural phenomena. Scientists can also *make* measurements, in which case the aim is to intervene in order to *produce* experimental objects and processes—characterized as 'effects'. (Keyser 2017, 2)

On Keyser's account 'taking a measurement' involves a scientist using a result in the context of theory to represent a given phenomenon (2017, 9-15). In contrast, 'making a measurement' involves setting up experimental conditions to produce a phenomenon—where that phenomenon can be realized in nature but it can also be a brand new phenomenon (Keyser 2017, 10). The difference between these two methodologies seems to be a matter of passive representation of a phenomenon vs. active intervention to produce a phenomenon. While the distinction between representation and intervention has been useful in classifying methodology in well-documented contexts like thermometry, microscopy, and cellular measurement, I argue that it falls apart in contexts where taking and making are *entangled*—such as in the context of biomarker measurement in the biomedical sciences.

In this discussion, I aim to show that in *complex methodological contexts*, representational and intervention-based roles require re-conceptualization. I analyze the *relations* between representation and intervention by focusing on the role of

intervention in *mediating* representations. In Section 2, I show how applied scientific practice challenges the simple distinction between representational and intervention-based roles of experiment/measurement. In Section 3, I discuss the complex interaction between representation and intervention applied to methodology in biomarker measurement.

2. Methodology at the Intersection between Intervention and Representation

In order to understand why the distinction between representation and intervention needs a multifaceted approach, it is important to be explicit about what it means to represent and intervene in scientific practice. In Section 2.1, I draw on van Fraassen (2008) to discuss representation and both van Fraassen (2008) and Keyser (2017) to discuss intervention. Then in Section 2.2, I show how applied scientific practice challenges the simplistic distinction between representational and intervention-based roles of experiment/measurement. I argue that the distinction between intervention and representation is less about *specific types of methodologies* in measurement/experiment and more about where one philosophically partitions the measurement *process*.

2.1. Representation and intervention

In experimental and measurement practice, representation has at least three important components: First, instruments or experimental contexts yield measurement values; Second, those values can only be interpreted within the context of a well-developed theory; and third, the relation between the measurement values and the phenomenon is determined by a user (e.g., experimenter). Van Fraassen (2008) provides

a rich characterization of representation in measurement and experiment, which requires careful analysis. Worth noting is that van Fraassen takes measurements to be a "special elements of the experimental procedure" (2008, 93-94). For my discussion the embeddedness of measurement in experiment is not important. I will focus on the roles or processes within measurement and experimental practice. But to do this, I will sometimes refer to 'measurement' and other times to 'experiment'. Van Fraassen's characterization focuses on interaction and representation in measurement:

> A measurement is a physical interaction, set up by agents, in a way that allows them to gather information. The outcome of a measurement provides a representation of the entity (object, event, process) measured, selectively, by displaying values of some physical parameters that—according to the theory governing this context—characterize that object. (2008, 179-180)

For van Fraassen, measurement interaction between an object of measurement and apparatus generates a physical outcome—the "measurement outcome" or "physical correlate of the measurement outcome"—, which provides information content about the target of measurement (2008, 143). The contents of measurement outcomes convey information about *what is measured* through the mediation of theory. Van Fraassen posits that theoretical characterization of measurement interaction requires 'coherence':

> The theoretical characterization of the measurement situations is required to be coherent with the claims about the existence of measurement outcomes, their relation to what is measured, and their function as sources of information. (2008, 145)

In short, the theory tells a coherence story about "how its outcomes provide information about what is being measured" (145). Furthermore, the information content is representational. Van Fraassen says, "The outcome provides a representation *of* the measured item, but also represents it *as* thus or so" (2008, 180). To understand how the representational relation works, it is important to refer to van Fraassen's 'representation criterion':

> The criterion for what sorts of interactions can be measurements will be, roughly speaking, that the outcome must represent the target in a certain fashion—, selectively resembling it at a certain level of abstraction, according to the theory—*it is a representation criterion*. (van Fraassen 2008, 141).

Two aspects of the representation criterion require explanation: First, the distinction between "target" and "outcome"; and second, the role of theory in the operation of measurement. I begin with the former. Van Fraassen makes a technical distinction between the target of measurement ('phenomena') and the outcome of measurement ('appearances'):

> Phenomena are observable, but their appearance, that is to say, *what they look like in given measurement or observation set-ups*, is to be distinguished from them as much as any person's appearance is to be distinguished from that person. (2008, 285)

For van Fraassen, phenomena are observable objects, events, and processes (2008, 283). He emphasizes that phenomena include all observable entities—whether observed or not (2008, 307). A given phenomenon can be measured in many different ways. The outcome of each measurement provides a perspective on a given phenomenon—meaning that the

content of measurement tells us what things *look like*, not what they *are like* (2008, 176, 182). The *content* of the measurement outcome is an appearance.

An important qualification is that for van Fraassen, a representation does not represent on its own. The scientist selects the aspects/respects and degrees to which a representation represents a target. This relation can be expressed as: Z uses X to represent Y as F, for purposes P.

Now that the target and outcome of measurement have been characterized, we can specify van Fraassen's role of theory in measurement. According to van Fraassen, "Measurement is an operation that locates an item (already classified as in the domain of a given theory) in a logical space, provided by the theory to represent a range of possible states or characteristics of such items (164). Three things are worth noting about van Fraassen's discussion of logical spaces. First, a logical space provides a multidimensional mathematical space that locates potential objects of measurement (2008, 164). By measuring we assign the item a location in a logical space. However, according to van Fraassen, it does not have to be on a real number continuum. As van Fraassen points out, items may be classified (by theory) on a range that is "an algebra", "lattice", or a "rudimentary poset" (2008, 172). Second, theoretical location depends on a "family of models" and not just an individual model (2008, 164). Third, an item is located in a "region" of logical space rather than at an exact point (2008, 165). Simply put, theory provides a classificatory system for what is measured. Importantly, theory is *necessary* for this type of classification. Van Fraassen says, "A claim of the form "This is an X-measurement of quantity M pertaining to S" makes sense *only* in a context where the

object measured is already classified as a system characterized by quantity M" (2008, 144 my emphasis).

We can summarize the above discussion into four conditions for van Fraassen's account of representation in measurement/experiment practice:

i. Physical Interaction Condition: The interaction between apparatus an object produces a physical correlate of the measurement outcome.

ii. Theoretical Characterization Condition: The content of the measurement outcome is given a location in a logical space, which is governed by a family of theoretical models. An item's location within a logical space can change in content and truth conditions as accepted theories change.

iii. Representational Content Condition: The content of a measurement outcome provides a selective representation of a given target of measurement (phenomenon). Because representations do not represent on their own, users and pragmatic considerations set the representational relation such that: Z uses X to represent Y as F, for purposes P.

iv. Perspectival Information Condition: Measurement generates appearances, which are public, intersubjective, contents of measurement outcomes. Appearances provide selective information about phenomena. Thus information from measurement tells us what something *looks* like and not what something *is* like.

Van Fraassen notes that measurement and experiment are not only limited to a representational role, they can take on at least two productive roles. First, instruments can produce phenomena that "imitate" natural phenomena. That is, carefully controlled conditions give rise to mimetic effects that are used by scientists in the context of theory to resemble natural phenomena (2008, 94-95). It is important to note that van Fraassen emphasizes that natural phenomena are phenomena that exist *independent of human intervention* (2008, 95). The second productive role of instruments is that they are used as "engines of creation" to produce or manufacture new phenomena. Van Fraassen is not explicit about whether or not the representational roles can smear with the productive roles. There is no reason to assume that these roles cannot be combined; but that requires explicit philosophical work to see *how*, which I develop in Section 3.

Keyser (2017) is explicit about the relationship between the representational and intervention-based roles in science. He discusses the *use* of intervention for developing causal representations. Scientists intervene, thereby manipulating causal conditions within a given measurement or experimental system, which he calls 'intervention systems', to produce some sort of "effect" (Keyser 2017, 9-10). According to Keyser, "Intervention systems consist of organized experimental conditions and as such the effects that emerge are often sensitive to changes in conditions" (Keyser 2017, 10). Once a given effect is produced it can be used in order to be informative about causal relations for theoretical model building.

Keyser (2017) also differentiates between the methodologies of taking measurements vs. making measurements. I interpret that taking measurements involves

three components: First, some instrument or experimental arrangement yields a qualitative or quantitative value; second, a 'theoretical representational framework'—which is just a body of models—is necessary in order to characterize that value according to parameters and relations between parameters; and third, a scientist sets up the resemblance relation between the measurement/experiment value and some aspect(s) of a phenomenon (Keyser 2017, 14-15). In contrast, when scientists make measurements they manipulate causal conditions—such as, preparatory, instrument, and background conditions—within an intervention system. This manipulation gives rise to some effect (Keyser 2017, 3-12).

There is something puzzling about Keyser's distinction between making vs. taking, if we apply the aforementioned conditions (i-iv): i. *Physical Interaction Condition*; ii. *Theoretical Characterization Condition*; iii. *Representational Content Condition*; and iv. *Perspectival Information Condition*. Namely, it seems that 'making measurements' is compatible with conditions i-iv, so it is not clear why there is a need for a distinction in methodological type, but rather just a difference in details for each condition. For example, when a measurement is made, there is a (i) *physical interaction* that occurs, but it is broader than just the instrument and object. The interaction can include "experimental conditions" (Keyser 2017, 3-5). The product of a made measurement is also amenable to (ii) *theoretical characterization*. Keyser emphasizes that theoretical characterization is necessary for experiment/measurement (Keyser 2017, 14); but he does not make the additional move to say that theoretical characterization is *part of the process* of making a measurement. That is, in order to make a measurement about an effect, one needs to also *characterize* that effect. Without the final

characterization, one is only dealing with the material conditions, which is an incomplete part of the measurement process. Keyser can accept that theoretical characterization is a necessary component of making a measurement. Otherwise, he risks offering a limited concept of 'making a measurement' that only applies to arranging the material components of the measurement process and nothing further.

The same challenge goes for (iii) *representational content* and (iv) *perspectival information*. An important component of the measurement process is to represent the relation between the produced effect and some aspect(s) of a phenomenon. For example, is this given effect a limited mimetic representation of a natural phenomenon or is it a brand new phenomenon? Without claims about what the effect is and its relation to objects, events, and processes in the world, 'making a measurement' is uninformative about part of the measurement process: the final value of the measurement outcome.

The aforementioned considerations question the need for a distinction between 'making' vs. 'taking'. One conclusion is that making uses the same components (i-iv), just with slightly different detail. But the other conclusion is a bit unsatisfying: making is really only about organizing the material components, which is an *initial* step in the measurement process, and it does not apply to later steps in measurement.

2.2. Dynamic relations between intervention and representation

I argue that the distinction between intervention vs. representation is less about *specific types of methodologies* in measurement/experiment and more about where to philosophically partition the *measurement process*. To make this point clear, I make two sub-points: 1) Measurement in the biological sciences offers complex and sometimes

blurred relations between instrument and object of measurement such that representation and production take on dynamic roles; 2) There is a difference between the act of measurement and the total process of measurement. I briefly describe (1) and (2).

On van Fraassen's (2008) and Keyser's (2017) characterizations of *representation* in measurement, the role of the instrument/apparatus seems to have an important mediating function. It may be the case that philosophical focus on case studies (e.g., thermometry, microscopy, cellular bio, and bacteria) that are instrument-intensive provide a certain support for an instrument-centric account of representation in measurement. Whether or not the necessary mediating role of instruments is an explicit part of both accounts, there is room to develop a richer philosophical view of the role of representation in the total measurement *process*. Without such philosophical development, we risk missing complex cases of measurement where intervention occurs side-by-side with representation. For example, in some cases of biological measurement, scientists use the organism to measure processes in that same organism but also to represent larger phenomena (Prasolova et al. 2006). For instance, mouse diets are manipulated in order to measure chromatin pattern changes. I characterize this as the mouse *constituting experimental conditions* that are being manipulated in order to measure some sort of process. The manipulation of conditions indicates an interventionist approach (or 'making' a measurement). Moreover, without manipulating the mouse's diet scientists would not be able to make a reliable measurement on chromatin structure at all. So the organism is not only being manipulated as part of the experimental/measurement set-up, it is a crucial part of that set-up. That is, without intervention, there is no reliable result. In addition to the organism being used as part of the measurement set-up, it also

serves as a physical *representation* of the dynamics of chromatin pattern change. That is, a given model organism can serve as a data model for a specific phenomenon of study—e.g., chromatin pattern in organism X. So, in this case the organism serves a dual function: it constitutes a set of experimental conditions to be manipulated and it serves as a physical representation of a phenomenon. Because of the dual function, this seems to be a case of both 'making' and 'taking' a measurement.

This brings me to sub-point (2). The total process of measurement is often complex in the biological sciences and requires multiple stages of intervening and representing. As mentioned in the model organism example representation and intervention are often *entangled*. Measurement is not merely putting an instrument up to something and waiting for a reading, which can be classified as an *act* of measurement. Measurement is also not merely creating effects out of material conditions. Measurement requires manipulation of conditions that is *used* in order to generate a representation. For example, identifying a mysterious fungus that is entangled with other fungus in a sample is an active process that requires both intervention and representation. One method is to take a sample and scrape it over a petri dish. What grows are spores that are passively deposited. But if common fungi were commingled with the mysterious fungi in the sample, and the common fungi grew faster, it would be impossible to identify the mysterious fungus. That is, coming back in a couple of weeks and seeing the petri dish covered with familiar species would lead to a false conclusion. Another way to perform the measurement (i.e. culture samples) is as follows. Take the samples and grind them up. Then sprinkle them into a petri dish. Put the dish under the microscope and, using a fine needle, pick out fragments of the mysterious fungus and transplant them to their own

dishes (Scott 2010). Once the fragments have been transplanted through this fine-grained intervention, each dish can be left to grow the colonies. The final dishes will offer visual representations that serve as data on the nature of the mysterious fungus. Notice here that intervention is a precursor to reliable representation.

Representation is not only reserved for the final instrument reading. It can also occur at other stages in the measurement process. Likewise, manipulation does not have to occur only at the earlier stages. For instance, organic matter can function as an instrument, like in the case of FourU thermometers, which are RNA molecules that act as thermometers in Salmonella (see Waldminghaus et al. 2007). Suppose that a scientist sets up an experiment to iteratively measure to what extent modifying RNA factors in FourU thermometers changes thermometer readings in Salmonella. In such a case the scientist could modify molecular factors and use the organic thermometers as temperature measures over many iterations, which would culminate in some sort of data model that organizes the relationship between molecular factors and FourU function. In such a case, there are multiple layers of intervention and representation.

The complex layering of intervention and representation is apparent in biomarker measurement in the biomedical sciences, where biological components serve as representations of disease conditions, but are also intervened on in order to make more reliable representations. I turn to this case study in the subsequent section.

3. Intervening in Representations and Representing Interventions

Biomarkers are used in biomedical measurement to reliably predict causal information about patient outcomes while minimizing the complexity of measurement,

resources, and invasiveness. A biomarker is an assayable metric—or simply, an indicator—that is used by scientists to draw conclusions about a biological process (De Gruttola et al. 2001). The greatest utility from biomarker measurement comes from their ability to help clinicians and researchers make conclusions with limited invasiveness. The reliance on biomarkers to make causal conclusions has prompted the use of 'surrogate markers'. These biomarkers are used to substitute for a clinically meaningful endpoint such as a disease condition. A major scientific methodological issue is that the use of multiple biomarkers will produce disagreeing results—and this is true even in the context of biomarkers that use similar biological pathways. To make methodological matters worse, theoretical representation is often not equipped to fill in the causal detail for each biomarker measurement. This amounts to an unfolding methodological puzzle about how to use intervention and representation in biomarkers to produce reliable measurements. My interest in this case study is not in solving the methodological puzzle, but rather in showing the *relations between intervention and representation* in such a complex case study. In this section, I discuss the complexity of intervention and representation in biomarker measurement to illustrate how intervention mediates the measurement process.

To understand the complex methodology in biomarker measurement it is important to detail the use and limitations of biomarkers. Some biomarkers are used as a substitute for some clinical endpoint. For instance, LDL cholesterol (LDL-C) is a biomarker that clinicians and physicians use to correspond to a clinical endpoint—e.g., heart attack. Moreover, the biomarker is associated with risk factors such as coronary artery stenosis, atherosclerosis, and angina pectoris. Katz (2004) argues that all biomarkers are candidates for 'surrogate markers', which can serve as substitutes for

clinical endpoints. That is, surrogate markers are reliable biomarkers that have a one-to-one correspondence with the disease condition such that they can be used to provide reliable predictive and causal information about a given clinical endpoint. There are a couple of points worth noting. First, notice that biomarkers and surrogate markers are being used as representations of a clinical endpoint. That is, to figure out the likelihood of developing a disease condition and to understand the risk factors associated with that disease condition, scientists use biomarkers that indicate information about the endpoint. This means that these physiological components can be used by clinicians and physicians to *represent disease conditions to respects and degrees*. The second point worth noting is that there are many biomarkers but limited surrogate markers and even more limited validated surrogate markers ('surrogate endpoints')—which are surrogate markers that are reliable in multiple contexts of interventions. The importance of this will be relevant shortly when I discuss the complexity of biomarker measurement. For our purposes, this means that most biomarkers in biomedical practice provide very limited representational information.

Surrogate markers are not passively used as physical representations of disease conditions. Their use is often more effective for representational purposes if there is a *mediating intervention*. For instance, surrogate markers can constitute "response variables". This is where a surrogate marker is manipulated in order to produce an effect that is relevantly similar to the effect with the same manipulation on the clinical endpoint. This means that an adequate surrogate must be "tightly correlated" with the true clinical endpoint; but it also means that any intervention on a surrogate marker must be tightly correlated with the intervention on the true clinical endpoint (Buyse et al. 2000). I

interpret this as a dual role for a reliable surrogate marker. It is to act as an epidemiological marker that *represents* some clinical endpoint but also to act as a responding variable that can be used in an *intervention* to causally influence the clinical endpoint. An example of the dual role of the surrogate marker is that high concentrations of LDL cholesterol (LDL-C) correspond to cardiovascular risk (Gofman and Lindgren 1950). But if a therapeutic interventions is used—such as, 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase inhibitors (statins)—that intervention can lower LDL levels, which in turn reduces cardiovascular disease (LaRosa et al. 2005).

So far I have presented the representational and intervention-based role of biomarkers. It is not straightforward to say that surrogate markers are *'made'* like an effect. But it is also not straightforward to say that surrogate markers constitute a *measurement outcome that is the final reading on an instrument*. These markers provide useful representational information *in the context* of an intervention. To add to the complexity of the relation between representation and intervention, biomarkers in the context of Alzheimer's measurement have added methodological steps. In Alzheimer's measurement there are different biomarkers, which are not correlated with each other and change with independent dynamics in the progression of Alzheimer's disease. So *each* of these biomarkers do not provide the same type of representation about the progression of Alzheimer's disease. Furthermore, scientists *only* understand the disagreement between each of these biomarkers in the presence of different interventions.[2] The different

[2] There has been much work recently on clinical biomarkers like: cerebrospinal fluid (CSF) tau, which is the primary component of neurofibrillary tangles; CSF 42-amino acid amyloid-β (CSF Aβ), which is the protein cleavage product believed to precipitate disease by forming neuron-damaging plaques; and amyloid plaques from PET scans.

interventions are in the form of drugs (e.g., bapineuzumab and solanezumab) and these interventions produce disagreeing representational results for the biomarkers. That is, the biomarkers respond differently to different interventions, which is methodologically problematic because it indicates that all of these biomarkers cannot be reliably tracking Alzheimer's progression in the same way. Interestingly, scientists systematically compare these disagreeing results to make reliable claims about Alzheimer's progression and treatment (Toyn 2015).[3] To simplify the method used, scientists track how interventions change properties of biomarkers and then they compare these amalgamated results with how interventions change behavioral/cognitive properties. This type of cross comparison allows scientists to eliminate biomarkers that do not track behavioral/cognitive improvement.

The structure of the methodological complexity in biomarker measurement can be partitioned as follows: 1) For a particular clinical endpoint, there are *limited physical representations* in the form biomarkers (or surrogate markers) which can be *used* to make representational and perspectival conclusions about the endpoint or risk factors associated with it; 2) *Scientists intervene in a process* from each of the biomarkers in order to track the relations between biomarkers and clinical endpoints; and 3) Such interventions

While the methodological story is beyond the scope of this discussion, there is a complex methodological point that is noteworthy for this discussion (Toyn 2015).

[3] To give a brief picture: The intervention of Bapineuzumab reduces levels of plaque assayed by Aβ PET and CSF tau, but not CSF Aβ; but Solanezumab *does not alter* levels of plaque assayed by Aβ PET and CSF tau but leads to a *reduction in* CSF Aβ. Cross comparison of the *intervention* mechanisms allows scientists to begin to make causal claims about which biomarkers are more reliable than others (Toyn 2015).

prompt *disagreeing results between the biomarkers*, which can 4) be amalgamated by researchers into further representations of the *relations between biomarkers and their clinical endpoints*. The above structural breakdown is merely *a* type of complex methodological process that can occur in biomedical measurement. It shows how interventions on physical representations (biomarkers) can produce other reliable representations. What is important to note about this analysis is the role of intervention in *mediating* further representations. In the case of biomarkers, intervention is necessary to test how close biomarkers are in their representations of clinical endpoints and also to other biomarkers. These representations not only represent the relation between the original biomarker and the clinical endpoint, but they also represent how a given intervention affects a given biomarker. As such, intervention paves the way for iterations of representations.

4. Concluding Remarks

In this discussion, I have analyzed the role of intervention in mediating representations by using examples from the biological and biomedical sciences. Characterizing intervention as a mediating factor in a larger methodological operation provides an important point about scientific practice. Representation and intervention are not neatly partitioned into contrasting methodologies. In fact, applied science often dictates the complex, and often smeared, philosophical concepts and methodologies. For this reason, I am proposing a *process* view of intervention and representation. This view opens up the diversity of relations between representation and intervention in a given experimental/measurement practice. While I have emphasized how intervention mediates

representation, there is more territory to explore about the mediating role of representation for intervention.

Work Cited

De Gruttola, V.G, Clax P, DeMets DL, et al. (2001). Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. Control Clin Trials 22:485–502.

Gofman, J.W., Jones, H.B., Lindgren, F.T., et al (1950). Blood lipids and human atherosclerosis. *Circulation* 2:161–178.

Hacking, I., (1983). *Representing and Intervening*, Cambridge: Cambridge University Press.

Heidelberger, M. (2003). Theory-ladenness and scientific instruments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 138–151). Pittsburgh, PA: University of Pittsburgh Press.

Katz, R. (2004). Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* 1:189–195. doi: 10.1602/neurorx.1.2.189

Keyser, V. (2017). Experimental Effects and Causal Representations. *Synthese*, SI: Modeling and Representation, pp. 1-32.

LaRosa, J.C., Grundy, S.M., Waters, D.D., et al. (2005). Intensive Lipid Lowering with Atorvastatin in Patients with Stable Coronary Disease. *New England Journal of Medicine* 352:1425–1435. doi: 10.1056/NEJMoa050461

Prasolova L.A., L.N. Trut, I.N. Os'kina, R.G. Gulevich, I.Z. Pliusnina, E.B. Vsevolodov,

I.F. Latypov. (2006). The effect of methyl-containing supplements during pregnancy on the phenotypic modification of offspring hair color in rats. *Genetika*, 42*(1)*, 78-83.

Radder, H. (2003). Technology and theory in experimental science. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 174–197). Pittsburgh, PA: University of Pittsburgh Press.

Toyn, J. (2015). What lessons can be learned from failed Alzheimer's disease trials? *Expert Rev Clin Pharmacol* 8:267–269. doi: 10.1586/17512433.2015.1034690

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.

Waldminghaus, T., Nadja H., Sabine B., and Franz N. (2007). FourU: A Novel Type of RNA Thermometer in Salmonella. *Molecular Microbiology* 65 (2): 413–24. https://doi.org/10.1111/j.1365-2958.2007.05794.x.

Respecting Public Investment: The Problems with Democratic Endorsement as a Criterion for Legitimate Value Influence in Science

Criticism of the value-free ideal has motivated attempts to formulate a criterion for the legitimacy of non-epistemic value influence in science. I argue that this search aims to protect two main components of legitimacy, scientific integrity and justice. While integrity is primary, justice remains important, especially in setting scientific goals. One of the main proposals for setting legitimate goals is to rely on democratic endorsement (Intemann 2015). I critically assess four interpretations of this criterion, finding that all are problematic. I then propose and evaluate three alternative models that seek to better balance respect for the public with scientific expertise.

I. Introduction

The results of scientific research inform practices quotidian to political: from what we eat for breakfast to how governments set policies to adapt to climate change. While the results of science are used in a variety of value-laden settings, the practice has often portrayed itself as objective and value-free.

Philosophers of science have heavily criticized the value-free ideal, arguing that non-epistemic value influence is unavoidable, sometimes even desirable, but also sometimes illegitimate (see Douglas 2016 for a review and references). The discussion has focused on determining a criterion for legitimate value influence.

In this paper, I explain two major approaches to this issue, showing how each requires a criterion for determining the legitimacy of scientific goals. I then discuss the notion of legitimacy, showing how it contains two main concepts: scientific integrity and justice. While I argue that integrity is more important, the question of justice remains. Next, I evaluate a prominent suggestion to determine legitimate goals for science: democratic endorsement. I propose four interpretations of democratic endorsement, showing how each is problematic to a different degree. Finally, I evaluate three alternative criteria that balance public dignity, public well-being, and scientific expertise.

II. Values in Science

In response to criticism of the value-free ideal, two approaches have been proposed to determine the legitimacy of non-epistemic value influence in science. The roles approach, developed most prominently by Heather Douglas, protects an epistemic core from influence

by non-epistemic values. Non-epistemic value influence may be legitimate only if it does not corrupt this core, sparingly defined as logical consistency and empirical adequacy. In this conception it is the ways in which values influence science, rather than the type of values, that determine legitimacy.

Douglas's framework makes two key distinctions (2016, 10). Value influence may be direct or indirect; the areas in which such influence occurs may be external or internal. External decisions occur outside of scientific study itself but may be about what to study or how to study it. Internal value influence can affect decisions about the methods of study such as which models to use or variables to measure.

Direct value influence is only acceptable when external, for example, when deciding which subjects to study. In contrast, indirect value influence is acceptable and sometimes desirable internally as well as externally. Indirect, internal, legitimate influence includes cases including inductive risk reasoning, which holds that evidential standards should respond to the consequences of being wrong. For example, it is rational to have high standards for the safety and efficacy of medicine, but it is also rational to have lower evidential standards for contamination in drinking water.

In contrast to the roles approach, the aims approach argues that value influence is legitimate if and only if it successfully serves legitimate goals. It is the kind of goals that determines the legitimacy of non-epistemic value influence, rather than the ways that such values may impact science. Kevin Elliott defines legitimate goals as ones that are transparent, representative, and inclusive of input of stakeholders (2017, 10). Kristen Intemann defines

legitimate goals as ones that are “democratically endorsed” (2015, 217). I will address the problem of how to clarify the concept of ‘democratically endorsed’ or ‘representative’ in section IV.

Both approaches balance the roles of epistemic integrity and legitimate goals for science in different ways. In the aims approach, the primary consideration is the legitimacy of goals, with the assumption that for value influence to successfully serve those goals it must protect epistemic integrity (Intemann 2015, 227). There are two ways this might occur: the protection of epistemic integrity may be a democratically endorsed goal, or epistemic integrity may be needed to achieve democratically endorsed goals. Either way, such protection is contingent on conditions that may not always be present. The ways approach, in contrast, explicitly protects epistemic integrity from non-epistemic value influence. However, external or indirect value influence is acceptable because it does not affect the epistemic core.

In the next section, I will argue first, that protecting epistemic integrity should take priority and second, that both schools still need to determine which kinds of values or goals are acceptable for the sciences.

III. Legitimacy as Justice and Integrity

One can understand the difference between the two approaches as a difference in their conception of “legitimacy.”

Douglas describes legitimate value influence in science as that which retains the trustworthy nature of science—a practice whose findings retain the epistemic integrity

normally expected of science. The definition is intuitively appealing because it attempts to protect the integrity of science as a practice. Without epistemic integrity, science fails to further the project of gathering reliable knowledge about the world.

The aims approach takes a different perspective on the notion of legitimacy, seeing it primarily as a question of justice. If legitimate value influence is that which successfully pursues legitimate goals, it is the goals of research that determine the legitimacy of such value influence. The goal of research cannot be simply to preserve epistemic integrity; it needs to serve some kind of justice.[1] This definition is appealing because it recognizes the societal power of science and the need to respect public investment in science. In addition, justice is an intrinsic good, arguably a characteristic to which all societal endeavors should aspire.

Integrity and justice, while both important, can come into conflict. I contend that integrity should take priority. Consider the four possibilities for non-epistemic value influence in science created by the two criteria: science that retains integrity and pursues just goals (J ∧ I), science that does not retain integrity but does pursue just goals (J ∧ ~I), science that retains integrity but pursues unjust goals (~J ∧ I), and science that neither retains integrity nor pursues just goals (~J ∧ ~I).[2]

1A thorough discussion of what counts as justice falls beyond the scope of this paper. We will consider two main components of 'justice': promoting public well-being and respecting public autonomy.

2 Here, just goals are either just or neutral, such as basic research, in contrast to unjust goals. The purpose of J is to exclude unjust goals.

1. J ∧ I. Science that pursues just goals and retains epistemic integrity is obviously desirable. For example, biologists might conduct careful study of the effects of hatchery salmon in order to ensure that local tribes are able to continue their traditional lifestyle.

2. J ∧ ~I. Science that pursues just goals but does not retain epistemic integrity is problematic. One could imagine an aquatic toxicologist who wanted to shut down a factory well-known for exploiting its under-age workers. She could choose methods of research and areas of observation that make the pollution of the factory seem worse than it is–for example, testing water directly next to the wastewater pipes from the factory and finding the concentration of lead higher than that allowed for river water. In this case, just goals are pursued—the factory is shut down—without the retention of scientific integrity. Her work does not provide fertile ground for future study or reveal anything about the world. It also risks undermining public trust in the scientific enterprise. Justice as a goal only makes sense when coupled with the pursuit of reliable knowledge.

3. ~J ∧ I. This form of science does not serve just goals, but does retain scientific integrity. For example, a scientist might be determined to show racial differences in intelligence. While her goals are deeply problematic, she does not allow bias to affect the epistemic integrity of her enterprise. This is also problematic: even if she continually fails to confirm her hypothesis, she uses resources on an endeavor that does not serve society's goals. At best, her failure lends support to the opposite view. At worst, the endeavor brings publicity to harmful viewpoints. Still, the scientist is producing useable (if not useful) data.

Both are valuable, but useful data serves important goals; usable data does not do so presently but is still valuable.

4. ~J ∧ ~I. The final category is science that serves unjust goals and does not retain epistemic integrity. For example, the racist scientist from the example above could manipulate her sample size until it "proves" her hypothesis. This is deeply problematic because, at worst, it promotes injustice and undermines public trust in science, and, at best, it is a complete waste of resources and fails to produce any kind of usable data.

We thus have four categories: J ∧ I (ideal), J ∧ ~I (well-meaning but problematic), ~J ∧ I (wasteful but producing useable data), and ~J ∧ ~I (highly problematic). It results from this analysis that integrity is more important. We should attempt to maximize both justice and integrity, but when in conflict, integrity should take priority. Consider ~J ∧ I and J ∧ ~I. Both use resources such as time, money, and public attention. ~J ∧ I uses these in pursuit of unjust goals, but produces usable data. J ∧ ~I, in contrast, may serve just goals, but in doing so produces non-useable data. Focusing solely on justice risks wasting resources *and* creating unusable data. Moreover, such science risks damaging the public trust that gives it privileged position in society.

Science is, basically, the building of a large self-correcting tower. J ∧ ~I introduces faulty bricks. Some bricks may be obviously faulty and easily discarded. But others may blend in with the rest, surreptitiously providing an inadequate foundation for other J ∧ I science. Integrity of construction comes first.

The aims approach fails to give sufficient priority to epistemic integrity because it focuses only on the goals of research and there are cases in which non-epistemic value influence can help science serve those goals successfully without retaining epistemic integrity. As discussed above, I and J are not coextensive; the assumption that successful science is science that retains its integrity is not well-founded.

Once the I/~I distinction can be made, however, the J/~J distinction remains extremely important. ~J ∧ I is still very undesirable because there are limited resources. J ∧ I is far preferable to ~J ∧ I. Even if integrity takes priority, we need a criterion for just goals in science. Both approaches to non-epistemic value influence in science, not only the aims approach, ought to articulate one. Let us turn to a critical evaluation of the main proposal in the literature.

III. Democratic Endorsement

Elliott (2017) and Intemann (2016) propose the following: science should serve democratically endorsed (or democratically representative) goals. We will first examine why this criterion is appealing before critically assessing four interpretations of the concept.

The criterion of democratic endorsement is attractive because the public has a large stake in the selection of scientific goals. The American public invests in scientific funding. The 2017 federal budget includes $155.8 billion for overall scientific research and development (Reardon & Ross 2017). Excluding Department of the Defense research, science spending comprises about 2.67% of federal discretionary spending. If science is publically funded, it is reasonable to ask that science be publically accountable—they are the

ones paying for it, after all. (Other major sources of scientific funding include research universities—many of which are publicly funded—and grants from private foundations.)

Additionally, the public shares in both the spoils and failures of science. Every day, people are saved by newly developed medical treatments; every day, people die from diseases without any available treatments. Science will be important as we enter an age of climate change that will be characterized by novel climatic and ecological conditions. The public bears the consequences of what science decides to study and not to study. In this way, the public also pays the price for the selection of scientific goals, even if such study is privately funded.

An argument for why democratic endorsement is attractive:

1. The public is invested in the outcomes of science.
2. When the public is invested in outcomes of an enterprise, their good should be considered in decision-making of that enterprise.
3. Therefore, the good of the public should be considered in scientific decision making.
4. The best way to accomplish 3 is to ensure the goals of science are democratically endorsed.

Premises 1-3 are relatively uncontroversial. Premise 4 is the primary assertion of Intemann and Elliott. It asserts that the good of the public is best served by democracy. (Premise 2 is a statement of justice: in this situation, just science acknowledges public investment in society by considering their good, defined as well-being and dignity.)

Democratic endorsement has an additional benefit of lending authority to science. The Declaration of Independence describes government as "deriving their just powers solely from the *consent of the governed*." Democratic endorsement might be rephrased as "consent of the affected," thus giving justification for goal selection.

Democratic endorsement as a criterion acknowledges public investment in science and lends science authority. However, it is unclear how democratic endorsement should be interpreted in practice. Below, I propose four possible interpretations of the criterion. Though some of these models are preferable to others, all are problematic.

The first interpretation is the simplest view of the term: "democratically endorsed" values are held by the majority of the polis. For example, a poll might show that 68% of the public values clean waterways, but only 45% value space exploration. On this interpretation, studying ways to restore riparian ecosystems would be legitimate, but space exploration would not be.

This simple view is both unacceptably permissive and unacceptably exclusive. First, the majority of the population might hold values that are problematic. Historically, the majority of the American population might have believed that black people were less deserving of rights than whites. This is the kind of value judgment that led to deeply problematic scientific practice like the Tuskegee syphilis study.

Second, the majority of the population might not hold a value that should influence science. For example, most scientists value environmental sustainability. However, public opinion might not reflect this value, whether today or historically. Research on the effects of

climate change is essential, and environmental sustainability seems to be an obviously legitimate value to drive that research. The exclusion of values such as environmental sustainability under this interpretation is problematic.

Additionally, the majority of the population would might not necessarily support the epistemic values that are essential to science. In a "post-truth" political environment that is normalizing "alternative facts," even truth may not reach the level of democratic representation necessary for legitimacy. While Intemann's and Elliott's work is focused on non-epistemic values, it is problematic that this conception could, when applied to epistemic values, exclude values such as empirical adequacy, a basic characteristic of scientific theories.

This interpretation illustrates a larger problem with democratic endorsement or representation: they give the public a responsibility to make decisions, even when it is not equipped to do so. Scientists possess expertise that shapes their values.[3] While one might worry about "undemocratically privileging" the values of scientists (Intemann 2015, 218), expertise should still be taken into account.

Let us turn to narrower interpretation of this model that sees democratically endorsed goals as ones that are commonly held: the consensus model. This model is less over-permissive than the simple-majority conception. Even if the majority of the populace holds prejudiced values, if at least some of the populace does not, then these values cannot legitimately influence scientific practice.

3 The confines of this paper preclude a full discussion of expertise. More can be found in Collins and Evans (2007).

However, the advantage of the consensus model is also a disadvantage: it is unclear that there exists sufficient common ground to provide an adequate pool of legitimate values. Even if the requirement is only that the vast majority agrees, very few non-epistemic values will meet such a standard. Several important non-epistemic values would most likely be disqualified, including environmental sustainability.

This model illustrates another general problem with democratic endorsement: the values of the populace are not temporally static. Recent paradigm shifts on opinions about LGBTQ issues illustrate the dangers of basing scientific practice on values that shift in popularity. Under this model, values might come in and out of consensus. The time scale of scientific research makes this problematic: early climate change research was not reflective of public consensus, but trends suggest that it already is or will be in the near future. If scientists waited for consensus to obtain, a wealth of scientific data on the climate would be missed. The consensus model improves over the simple majority model, but shares the problem of excessive exclusivity.

Values exist on different levels of specificity. A different version of the consensus model that focuses on underlying values could be more successful.

As an illustration, consider two groups. One values the continued existence of mountain ecosystems in West Virginia, because, for example, they care about their children's ability to have certain recreational experiences. Another group might value the continued existence of coal mining jobs, for example, so as to preserve their economic benefits. Under

the simple model, whichever group is the largest would have the legitimate value. Under the consensus model, neither value would likely be legitimate.

Now consider the underlying values. Those who value mountains probably do so because of a more fundamental value of environmental sustainability. Likewise, those that value coal jobs probably do so because of a more fundamental value of economic sustainability. Looking one step deeper, we find that environmental sustainability and economic sustainability both focus on sustainability tout court: the ability to continue our existence. Sustainability might flow from values like prudence or a hope to provide for future generations. Interests that conflict at the surface level may share common underlying values.

The foundational-consensus model improves over the simple or consensus model by looking at the shared values that inform more specific ones, thus expanding common ground. The consensus model would exclude both environmental and economic sustainability, but the foundational-consensus model would allow the underlying value of prudence.

This model addresses the first general problem because it provides scientists with foundational values to interpret with expertise. It allows some public influence as well as room for scientific discernment.

While foundational values are likely more stable, change is still possible. The emphasis on certain values may shift. For example, in the 1950's, progress and growth probably qualified as foundational values. The same might not be true in 2040.

Additionally, by looking beyond the specific, the model might have become too vague. If very different specific values can be supported by the same foundational ones, it

follows that these foundational values provide little guidance to set specific goals. For example, scientists may decide to interpret the foundational value of prudence as environmental sustainability and work on climate change. That decision might contradict the specific values of a population who views prudence differently. It becomes unclear whether the scientists' goal is democratically representative. If considering foundational values can lead to drastically different goals, the model seems unhelpful.

The foundational-consensus model might be improved if we look at fundamental values through a different lens, possibly in a way similar to John Rawls's veil of ignorance (1971/1999, 118). In the thought experiment, Rawls asks us to imagine ourselves as rational individuals planning a society from a standpoint where we do not know anything about our position in society. Rawls argues a society thus planned would be based on his well-known principles of justice. The veil of ignorance is meant to make us think outside of privilege, making justice a part of rational self-interest.

Could we set appropriate, democratically endorsed, goals for science from behind the veil of ignorance? On the one hand, some appropriate epistemic goals would likely be endorsed: rational individuals in the original position would likely want science to be successful and see its spoils justly distributed. On the other hand, problems arise when we try to interpret the idea that scientists ought to "check values" at the door. The entire motivation behind the pursuit of our criterion is the recognition that we cannot separate ourselves from our values; a thought experiment that relies upon that ability is unlikely to help.

The intuitive appeal of Rawls's veil of ignorance points to the idea that science ought to serve just goals. No one person's values should be privileged without sufficiently compelling reason.

Are there compelling reasons to privilege the values of scientists? In some situations, yes. Scientists have epistemic privilege from their expertise that informs their values. While democratic endorsement is initially attractive, it is problematic because not enough room is left for the importance of scientific expertise.

IV. Other Possible Criteria

We can now move onto our next project of investigating whether there is a better criterion to determine the legitimacy of non-epistemic value influence on scientific goals.

In order to do so, we need to ask what we want in such a criterion. If we return to the syllogism presented in Section III, premises 1-3 stand. It is premise 4—that democratic endorsement is the best way to respond to public investment in science—that is problematic. Our task is to better fulfill premise 3–that public interest be considered.

Democratic endorsement does not successfully fulfill premise 3 because it fails to take scientific expertise seriously. Expertise informs values and goals; scientists have epistemic privilege not available to the public that should be considered.

Thus, a desideratum is that scientific expertise is taken seriously. Another is that the public investment be respected. Such respect can manifest in two ways: through the respect of public dignity and the respect of public well-being, depending on whether one focuses on deontology or utility.

So, three elements ought to be considered: public well-being, public dignity, and scientific expertise. I propose three models to be assessed: paternalism, professionalism, and the trustee model. The models differ on the issues of the presence and justification of authority.

Before we evaluate these models, it is important to discuss how the three elements above relate to one another. It is inaccurate and unhelpful to see them as ends of one-dimensional spectra with high respect for public on one end opposed to high respect for scientific expertise on the other. I propose to see them as different axes on a three dimensional space. One axis represents degrees of respect for the public, the second represents degrees of respect for scientific expertise, and the last represents degrees of respect for the public well-being. Thus, it is a priori possible to maximize all elements.

With this in mind, let us turn to our assessment of our three potential models. The first is paternalism, where the values and decisions of those with epistemic privilege are favored and imposed on those viewed as less capable or incapable of making those decisions, without consent (Dworkin 2013). Paternalism is justified (if ever) only if it is motivated by the well-being of its target. The classic example of paternalism is of a parent and child; it is not just for a parent to treat her toddler's wishes about the best way to cross the street equally to her own. A paternalistic model allows scientists to set goals for science on the basis of their expertise and with the public's well-being in view.

Paternalism exhibits high respect for scientific expertise, but low respect for public dignity. This model could promote well-being if we agree to the premise that the epistemic

privilege of scientists gives them an advantage of knowing what is best for the public (an admittedly controversial claim). Paternalism is justified only in specific conditions where it is acceptable that the public is deemed incapable of making appropriately informed decisions. That such conditions be ever satisfied is disputable.

In view of our assessment of paternalism, let us look for a model grounded in public consent: the professional model (Freidson 2001, 180). In a market economy, the proper role of businesses is not to tell customers what they should want. It is not the place of the car dealer to tell a customer who comes to buy a two-door sports car as her family's main vehicle that a minivan would serve her four children better. This contrasts with the role of professionals. Educators are accountable to society as a greater whole and tell the students what they should want, even if that prescription is a long term paper. Professionals serve a higher good than the immediate preferences of the person they are serving.

This model diverges from paternalism in that one is free to take or leave the advice offered by professionals. Doctors provide professional advice, but their patients are not forced to follow suit. The professional model, applied to scientific goal-setting, construes scientists as providing advice that the public may or may not follow. Thus there is high respect for public dignity, but lower emphasis on scientific expertise. There is a potential that the public will choose against the informed opinion of scientists, and so potentially negatively affect their own well-being. Professionalism overly privileges public dignity at the expense of the other desiderata.

Is there a middle option? Political philosophy provides a distinction between the trustee and representative models (McCrone & Kuklinski 1979). In the trustee model, the expert has increased access to relevant knowledge and is entrusted with decision-making powers by a group. The trustee serves the group's best interest, even if her decision does not reflect those of the group. A classic example is an investment banker who acts in the fiduciary benefit of her clients. The investors do not pay the banker to make the decisions they themselves might make; rather, they pay her to use her privileged knowledge to make decisions that best serve their interest. In contrast, the representative model sees the decision maker as a proxy for the will of the people. Her job is primarily to represent the decisions that they might have made, sometimes despite their best interest. Scientists seem to fit better in the trustee model than they do in the representative model. Like investment bankers who know the market better than their clients, scientists know the scientific landscape better than the public.

This trustee model respects scientific expertise. It also respects public well-being, under the premise that scientific expertise is a good guide to determine it. Finally, it respects public dignity, as it is grounded on consent and trust. The trustee model thus fulfills our desiderata. That said, two difficulties arise. First, while the best interest of the investment bankers' clients is clear, the public interest in science is less so. Second, while many people choose to sign contracts with bankers, it is not clear that the current public would willingly entrust scientists with decision-making power over the goals of science

The trustee model best maximizes our three desiderata, but it is unclear whether the conditions under which it is justified currently exist. In its absence, the choice of which model is preferable between paternalism and professionalism is unclear. Paternalism sacrifices public dignity for well-being; professionalism lowers the ability of scientists to promote well-being of the public, but gives the public greater autonomy. We are left with a classic moral dilemma: do we value public well-being or respect for public dignity? Kant says we should promote autonomy above all; he would prefer the professional model. Mill, in valuing well-being, would likely be more sympathetic to the paternalism model. Among the three models, we are left with an, appropriately value-laden, choice. Of course, an obviously desirable fourth option is a different model than those proposed in this paper that better maximizes our three desiderata.

V. Conclusion

If we accept that non-epistemic value influence in science is inevitable, we need a way to determine when such value influence is legitimate. Legitimacy has two components: retention of epistemic integrity and promotion of justice. Even if democratic endorsement is asked to do only the work of promoting justice, it fails to do so because it fails to take scientific expertise seriously. We need a new model in which public dignity, public well-being, and scientific expertise are respected. We assessed three candidates: paternalism, professionalism, and the trustee model. While the trustee model seems ideal, its conditions of realization may not be satisfied in our society. If so, we are bound to choose between professionalism, a more deontological approach, and paternalism, a more utilitarian

approach, for goal-setting in the sciences.

References:

Collins, Harry and Robert Evans. *Rethinking Expertise*. University of Chicago Press, 2007.

Douglas, Heather. 2016. "Values in Science." In *Oxford Handbook of Philosophy of Science,* ed. Paul Humphries. Oxford University Press: 609-30.

Dworkin, Gerald. 2013. "Defining Paternalism." In *Paternalism: Theory and Practice,* ed. Christian Coons and Michael Weber. Cambridge University Press: 25–39.

Elliott, Kevin. 2017. *A Tapestry of Values*. New York: Oxford University Press.

Freidson, Eliot. 2001. *Professionalism, the Third Logic.* University of Chicago Press.

Intemann, Kristen. 2015. "Distinguishing Between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5 (2): 217-32.

McCrone, Donald and James Kuklinski. 1979. "The Delegate Theory of Representation." *American Journal of Political Science* 23 (2): 278-300.

Rawls, John. 1971/1999. *A Theory of Justice.* Cambridge, MA: Harvard University Press.

Reardon, Sarah and Erin Ross. 2017. "Science Wins Reprieve in US Budget Deal." *Nature News.* doi:10.1038/nature.2017.21835.

Rooney, Phillis. 1992. "Is the Epistemic/Non-Epistemic Distinction Useful?" *Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992: 13-22.

Function Words and Context Variability*

Shane Steinert-Threlkeld
S.N.M.Steinert-Threlkeld@uva.nl

Draft of 30 October 2018. Comments Welcome!

> *Twas* brillig, *and the* slithy toves
> *Did* gyre *and* gimble *in the* wabe;
> *All* mimsy *were the* borogoves,
> *And the* mome raths outgrabe.
>
> Excerpt from 'Jabberwocky', in Carroll (1871), emphasis mine.

The poem excerpted in the epigraph has often been called a 'nonsense poem'. After all: what does it mean? What is a slithy tove? What does it mean to be brillig or mimsy? Calling it nonsense, however, overlooks the amount of meaning we can extract from the emphasized words: minimally, a scene in the past is being described, which took place somewhere called a 'wabe'. The emphasized words are what are known as *function words*: they provide the 'grammatical glue' among the *content words*, which are indeed nonsense in this excerpt.

The distinction between these two types of expression occupies a central place in modern theoretical linguistics. Rightfully so: every natural language exhibits a distinction between function and content words. Yet surprisingly little has been said about the emergence of this universal architectural feature of natural languages. Why have human languages evolved to exhibit this division of labor between content and function words? How could such a distinction have emerged in the first place?

This paper takes steps towards answering these questions by presenting a simple model of trial-and-error language learning in which a division of signals into function and content words emerges. In the next section, I briefly but more explicitly introduce the distinction. In Section 2, I argue that a necessary condition for the emergence of the distinction is the presence of *non-trivial composition* (in a sense to be made precise). I present three case studies in which only trivial composition emerges and a mathematical result that diagnoses why that is the case. In Section 3, I introduce a new type of signaling game – the Extremity Game – in which the objects of communication vary from play to play. Amidst such variation, a distinction between function and content words could be useful. Section 4 reports an

*Acknowledgments to be added. This work was supported by funding from the European Research Council under the European Unions Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

experiment, in which artificial neural networks are trained by reinforcement learning to communicate in the Extremity Game. The emerging languages are analyzed: when the agents can pay attention to perceptually salient features of the context, they learn a system with complex signals that we can interpret as a gradable adjective plus a superlative morpheme (a prime example of a functional item). Section 5 concludes.

1 Functional and Lexical Categories

Modern theoretical syntax distinguishes between two broad types of syntactic categories: lexical and functional.[1] The former broadly correspond to the major parts of speech: nouns, verbs, adjectives, and adverbs. The latter are a bit more varied, but include:[2]

- Prepositions: 'in', 'above', 'from', 'to', ...
- Determiners: 'a', 'the', 'every', 'some', 'many', ...
- Conjunctions: 'and', 'or', ...
- Complementizers: 'that', 'if', 'whether', ...
- Tense:
  - Auxiliaries: 'have', 'is', 'was', ...
  - Modals: 'will', 'would', 'can', 'might', 'ought', ...

Exactly characterizing the distinction remains tricky. After observing that lexical categories have 'contentful' meaning, while functional categories have 'grammatical' meaning,[3] it is usually observed that the former constitute *open classes* and the latter *closed classes*. Roughly: one can very readily introduce new nouns and verbs to a language as needed. By contrast, trying to introduce a new preposition 'below' meaning partially above and partially below would be quite difficult. Kaplan (1978) famously tried to introduce a new expression 'dthat', which rigidly referred to the satisfier of a description. Though he ably demonstrated the use of such a tool, that it never caught on can be partially attributed to the fact that demonstratives belong to a closed class. For the purposes of this paper, these distinctions suffice to point to the intended contrast.

Before proceeding, it's worth highlighting that the field of semantics – the scientific study of linguistic meaning – roughly divides itself along the lexical/functional line as well. The tradition descending from Montague via Partee and many others, usually called formal semantics, studies specifically compositional semantics. A survey of the textbooks in this field[4] shows that the major expressions studied come exactly from the functional categories.

[1] See, for example, pp. 43-46 of the textbook Carnie (2006).

[2] This list is incomplete and meant to be illustrative only. There are some debates about exactly which category certain expressions belong to, but they are orthogonal to present concerns.

[3] See, e.g., Carnie (2006) and Rizzi and Cinque (2016). Muysken (2008) is a thorough overview of functional categories.

[4] For example, Heim and Kratzer (1998) and Jacobson (2014).

Lexical semantics – the study of the meanings of basic expressions – studies at length the meanings of individual expressions and groups thereof in the lexical categories.[5] Seen in this light, explaining the emergence of the distinction between functional and lexical categories occupies a central role in the broader explanation of the emergence of compositionality.

2 Non-Trivial Composition

In this section, I build on the foregoing remarks in order to argue for the following claim: for a communication system to have function words, there must exist *non-trivial composition* (in a sense to be made precise) of complex signals. After presenting this argument, I will analyze three case studies from the literature on the evolution of compositionality which exhibit only trivial composition. The reasons for this are then made precise in the form of a triviality result: given the assumptions about optimal communication often made, the resulting systems must be trivially compositional.

The principle of compositionality says that the meaning of a complex expression is determined by the meanings of the parts and how they are put together.[6] Natural languages are compositional: whence the ability of competent speakers to produce and comprehend a potentially infinite set of novel expressions. A language can, however, be compositional without exhibiting the rich flexibility that human languages do. We will use the following definition:[7]

(1) A communication system is *trivially compositional* just in case complex expressions are always interpreted by intersection (generalized conjunction) of the meanings of the parts of the expression.

The force of this definition can be brought out by an example: Titi monkey calls.[8] In a series of predator-model experiments, it was found that raptors in the canopy elicit sequences of A calls, cats on the ground elicit sequences of B calls, cats in the canopy elicit one A followed by a sequence of Bs, and raptors in the canopy elicit a sequence of As followed by a sequence of Bs. While the full details do not concern us,[9] Schlenker, Chemla, Schel, et al. (2016a) argue that the best analysis of this call system involves the following semantics, interacting with some plausible pragmatic principles:

(2) Compositional semantics of Titi alarm calls: where t is a time,
   a. $[\![B]\!]^t = 1$ iff there is a noteworthy event at t
   b. $[\![A]\!]^t = 1$ iff there is a serious non-ground alert at t
   c. $[\![wS]\!]^t = 1$ iff $[\![w]\!]^t = 1$ and $[\![S]\!]^{t+1} = 1$
      [where w is a call and S a sequence of calls]

The crucial feature of this semantics concerns the rule (2c) for interpreting complex expressions (sequences of calls). It says that a sequence of calls is interpreted by first evaluating

[5] See, for example, Levin and Rappaport Hovav (2005).
[6] Frege (1923), Janssen (1997), Pagin and Westerståhl (2010a), and Pagin and Westerståhl (2010b).
[7] For this use, see Schlenker, Chemla, Schel, et al. (2016b) and Zuberbühler (2018).
[8] Cäsar et al. (2013) and Schlenker, Chemla, Schel, et al. (2016a).
[9] See Steinert-Threlkeld (2016b) for some reservations about the full analysis.

the beginning of the sequence at time t, then evaluating the rest of the sequence at time $t+1$, and conjoining the results. This clause results in the following: each call in the sequence contributes to the meaning of the whole *independently* of the other calls, with the complete meaning resulting from conjunction. It thus constitutes a paradigm of the definition of trivial compositionality in (1).[10]

In other words, non-trivial compositionality involves non-conjunctive modification of one linguistic item by another. Examples of such systems can also be found in communication systems much simpler than human language. In particular, Campbell's monkeys have been argued to exhibit it.[11] They have two basic alarm calls: an eagle call *hok* and a general alert *krak*.[12] Moreover, both calls combine with what appears to be a suffix *-oo*, which has the effect of weakening the severity of the calls. Schlenker, Chemla, Schel, et al. (2016a) propose the following semantics:

(3) $[\![R\text{-}oo]\!]^t = 1$ iff at t the sender is alert to a disturbance that licenses R but that is not strong among such disturbances.

This is non-trivial: *-oo* does not contribute independent meaning that is then conjoined with the contribution of *hok* or *krak*. Rather, it combines with one of the latter calls to modify the normal meaning of that call.

Here is the simple argument for the claim that non-trivial composition is necessary for the emergence of function words. Recall the characterization thereof as 'grammatical glue': they precisely do not contribute independent content to a sentence, but structure that provided by the content words. In a trivially compositional communication system, each expression contributes independent meaning to the complex expressions containing it. Therefore, none of the expressions therein are function words.

Before proceeding, we note that the presence of non-trivial composition does not suffice for the presence of function words. To see this, consider subsective adjectives.[13] These are adjectives like 'skillful', which have the property that for every noun, a 'skillful N' is an N, but is not 'skillful' in any sense independent from the noun. For example:

(4) a. Jakub is a skillful rock climber.
b. Jakub is a cook.
c. Therefore, Jakub is a skillful cook.

The inference pattern in (4) is not valid: Jakub can be skillful at one thing but not at another. If 'skillful' contributed its meaning independently of the noun it combines with, the inference would be valid: Jakub would be a climber, a cook, and skillful; therefore, a skillful cook. But 'skillful' is still a content word. One could imagine a very simple language whose only complex expressions were of the form 'Adj N', but which had subsective adjectives. This language would be non-trivially compositional but would have no function words.

[10] Berthet et al. (2018) argue that the proper semantics for Titi calls is not in fact trivially compositional. Nevertheless, the presentation just given illustrates what such a system would look like.

[11] Ouattara, Lemasson, and Zuberbühler (2009) and Schlenker, Chemla, Arnold, et al. (2014).

[12] The possibly different meaning of *krak* in different habitats of Campbell's monkeys is the subject of the aforementioned papers. We follow Schlenker, Chemla, Schel, et al. (2016a) in giving it a general meaning.

[13] Partee (1995).

Now, I will present three case studies of prominent models purporting to explain aspects of the evolution of compositional communication. Each of them, however, will turn out to exhibit only trivial composition. After presenting the case studies, I identify common underlying assumptions and then prove a mathematical fact demonstrating that under those assumptions, the resulting communication systems must be trivially compositional. In light of the foregoing, none of these extant approaches can explain the emergence of the distinction between function and content words.

2.1 Three Études

Nowak and Krakauer (1999) apply mathematical models of natural selection to the evolution of language, providing conditions under which a 'grammatical' language will evolve from a non-compositional one. In their model, states are object-action pairs, loosely modeling events. They compare two types of languages: one in which each object-action pair has an independent label, and another in which each object has a corresponding expression, each action has a corresponding expression, and the agents communicate by sending the corresponding pair of expressions to communicate about an object-action pair. While the results they obtain are indeed interesting, it should be clear from this brief exposition that the type of language that they consider exhibits only trivial composition: each component of a complex expression contributes its bit of meaning (either an object or an action) independently of the other.

Barrett (2007) and Barrett (2009) studies a generalization of signaling games[14] with multiple senders. In the simplest case, there are four states of nature and two sender, each of whom can send one of two signals to one receiver. The senders, but not the receiver, know which state obtains. Simulations show that a simple form of reinforcement learning leads these agents to a situation of perfect communication. Given the nature of the setup, the resulting systems look as follows. One sender partitions the four states into two sets of two, one for each signal. The other sender sends its two signals in an *orthogonal* partition.[15] One can imagine the states as a two-by-two square, with one sender indicating the row and the other the column of the true state. Such a system again exhibits only trivial composition, since the meaning of each sender's signal is independent of the other's and the receiver interprets the sequence by intersecting the two.

Finally, Mordatch and Abbeel (2018) study the emergence of communication in a multi-agent setting where each agent has a private goal that it wants to achieve.[16] The agents – which are in this case recurrent neural networks – communicate about a world with various colored landmarks in it. Each agent additionally has a color and its own perspective from its position (i.e. no agents share a frame of reference). The goals consist of getting an agent to perform an action (going to or looking at) at one of the landmarks. With appropriate costs for maintaining large lexicons, the agents learn to send sequences of signals with separate signals for which agent, which action, and which landmark. These three types of signals have independent meanings, which are combined by conjunction.

[14] Lewis (1969) and Skyrms (2010).
[15] See, e.g., Lewis (1988).
[16] The set of goals is assumed to be consistent, i.e. all of the goals are simultaneously realizable.

2.2 A Limitative Result

There is in fact an underlying reason that these systems exhibit only trivial composition. Although the three cases just illustrated come from different theoretical frameworks, they all share the same following assumptions:

(A1) Agents communicate about a fixed set of states. (Object/action pairs, separate points of a state space, and agent/landmark/action tuples, respectively.)

(A2) Optimal communication consists in correctly identifying the true member of the state space.

(A3) Messages are fixed-length sequences of signals from fixed sets.

It turns out that under these assumptions, there's a mathematical sense in which optimal communication will be trivially compositional. This is captured in the following result:

(5) Let X and $\{M_i\}_{i\in I}$ be any sets, and f, g two functions of the following type:

$$X \xrightarrow{f} \prod_i M_i \xrightarrow{g} X$$

Define $f_i^{-1}(\overrightarrow{m}) := \{x \in X : f(x)_i = \overrightarrow{m}_i\}$. Then the following holds.

$$\text{If } g \circ f = \mathrm{id}_X, \text{ then for all } \overrightarrow{m}, \{g(\overrightarrow{m})\} = \bigcap_i f_i^{-1}(\overrightarrow{m})$$ [17]

Here, X represents the fixed set of states about which the agents communicate. Note that the structure of this set does not matter. $\prod_i M_i$ is the set of possible sequences of signals, with each M_i being the signals available to be sent in position i of a sequence. f is a sender function: a function from states to sequences of signals. This can capture a single sender, or multiple acting either independently or in concert. g is a receiver function: it decodes the sequence of signals to one of the states X. Because id_X is the identity function on X, mapping each point to itself, that $g \circ f = \mathrm{id}_X$ means that optimal communication has been achieved, in the sense that the receiver always recovers the true state from X. Under that assumption, the result says that the receiver interprets a complex message (a sequence) by *intersecting* the independent meanings of each signal in the sequence (represented by $f_i^{-1}(\overrightarrow{m})$).

This result identifies three assumptions that cannot all be maintained if one wants to model the emergence of non-trivial composition, which I have just argued is a necessary step for explaining the emergence of function words. Not every approach makes all three of these assumptions. In particular, Steinert-Threlkeld (2014) and Steinert-Threlkeld (2016a) as well as Barrett, Skyrms, and Cochran (2018) drop (A3). In these models, not every message is a

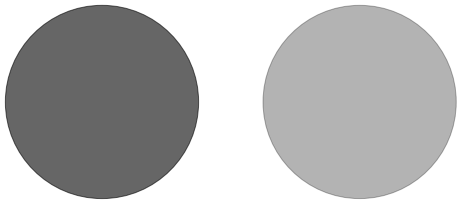
[17] *Proof:* Note first that g must be a surjection and f an injection. Without the former, there would be an $x \in X$ that is not $g(\overrightarrow{m})$ for any $\overrightarrow{m}$, and so $g \circ f \neq \mathrm{id}_X$. Without the latter, distinct points in X would get mapped to the same point in X by $g \circ f$. Now, suppose there were an $\overrightarrow{m}$ such that $\{g(\overrightarrow{m})\} \neq \bigcap_i f_i^{-1}(\overrightarrow{m})$. This can hold only if $\bigcap_i f_i^{-1}(\overrightarrow{m})$ contains more than one element, since $g(\overrightarrow{m})$ has to belong to the intersection. This entails that there is another point $x \neq g(\overrightarrow{m})$ for which $f(x) = \overrightarrow{m}$, contradicting the injectivity of f. □

sequence of the same length. In the former, one sender can choose whether or not to prefix a set of signals with an additional signal. In the latter, two senders choose *whether or not* to send a signal, so messages can be either of length one or two. In either case, the message space is a union, not a product (i.e. not of the form $\prod_i M_i$ for any sets M_i), and so the limitative result does not apply.

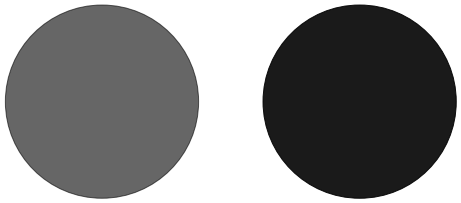
In the remainder, I will develop a model which maintains (A2) and (A3) but drops the assumption (A1) of a *fixed* set of states that the agents communicate about. That is: the context in which the agents are communicating will vary. Against that backdrop, there will be a role for function words to play.

3 A Signaling Game with Varying Contexts

The variant on the signaling game that I will use to illustrate the emergence of function words will have the agents talking about varying sets of objects with multiple *gradable properties*. To get a feel for the kind of task involved, consider the following adaptation of an example from Graff (2000).[18] Suppose that we are both looking at the following two circles, drawn on top of a table.



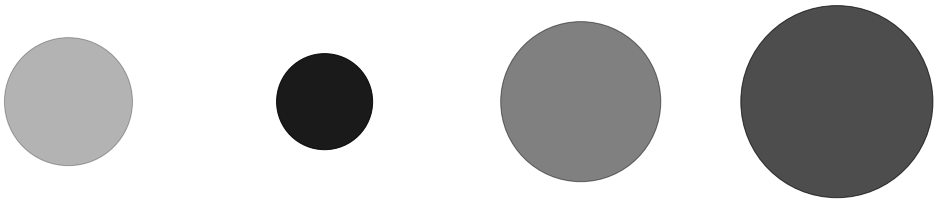
For whatever reason, you need me to put something on the left circle. You might say "put it on the *darker circle*". By contrast, suppose that you had the same communicative needs, but now the circles on the table looked as follows.



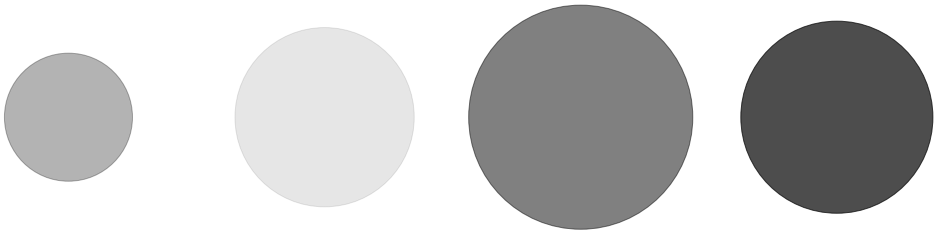
Now, to tell me to put it on the left circle, you might say "put it on the *lighter circle*", or a bit more circuitously "put it on the less dark one".

The target referent of your communication – the circle on the left – has exactly the same size and shade in both contexts. But in one context, it's dark*er* than the other circle, while in the other context, it's light*er*. (For the purposes of illustration, we can assume that you could not refer to the circles by their spatial position or demonstratively. If you'd like: your friend is looking at a picture on a screen that may have been scrambled.) Finally, we can imagine similar situations with more than one gradable property. Suppose, again, that you need to communicate about the leftmost circle in the following array.

[18] See Syrett, Kennedy, and Lidz (2010) for a study using similar contexts with children.



Here, it's natural to call the leftmost circle "the lightest one". Now consider the following context.



Here, you are likely to refer to the circle on the left as "the smallest one". These situations have the following structure: in each context, each object has two very salient gradable properties: a size (radius) and a darkness. These dimensions distinguish the target object: it has either the largest or the smallest value in one of those dimensions. By drawing attention to that fact, one can successfully refer to it. Moreover, you can do so in a very economical way: with labels for the properties and morphemes like the superlative *-est* (and its corresponding negative counterpart, 'the least'), successful communication is ensured. This is done without talking about specific degrees of size or of lightness and in a way in which an object with exactly the same degrees on all relevant properties will be referred to in different ways in different contexts.

I will convert communicative scenarios like the above into a type of signaling game – called the Extremity Game – with a few helper definitions. Following the literature on gradable adjectives,[19] I will assume that objects have some number of gradable properties, where each property has a corresponding *scale*. A scale in turn is a set of *degrees*, totally ordered with respect to a dimension. For example, the size of a circle corresponds to its radius, with degrees being positive real numbers (i.e. $\mathbb{R}^+$). For the degree of an object o on a scale s, I will write $s(o)$. Given a set S of scales, I will define a context as follows.

(6) A *context* c over scales S is a set of objects such that: for each $o \in c$, there is a scale $s \in S$ such that either o has the least degree on s ($o = \arg\min_{o' \in c} s(o')$) or the highest degree on s ($o = \arg\max_{o' \in c} s(o')$).

At its most general form, the game takes place between a sender and a receiver in the following way.

(7) Extremity Game, in general:
   a. Nature chooses a context c and a target object $o \in c$.
   b. The sender sees c and o and sends a message m from some set of messages M.

[19] See, for instance, Kennedy and McNally (2005) and Kennedy (2007) and the references therein.

c. The receiver sees c and m and chooses an object o' from c.

d. The play is successful (and the two agents equally rewarded) if and only if $o' = o$.

To fully specify a game, one must say what the messages M available are and how the agents make their choices. I will specify the former now and the latter in the next section. The set of available messages will be inspired by the semantics for gradable adjectives. There, it is assumed that adjectives map objects (of type e) on to their degree on the corresponding scale (of type d). Morphemes like *-est* and *least* then map a contextually specified set of objects to the subset with the highest and lowest degrees.

(8) Toy semantics for a gradable adjective and superlative morphemes.

a. $[\![\text{size}]\!] = \lambda x.s_{\text{size}}(x)$

b. $[\![\text{-est}]\!]^c = \lambda P_{\langle e,d\rangle}.\lambda x_e.x \in c \text{ and } \forall x' \in c, P(x) \succeq P(x')$

c. $[\![\text{least}]\!]^c = \lambda P_{\langle e,d\rangle}.\lambda x_e.x \in c \text{ and } \forall x' \in c, P(x) \preceq P(x')$

Now, for the crucial observation: in contexts as defined in (6), having one expression for each scale and the morphemes *-est* and *least* will suffice to uniquely pick out each object in the context. I will assume, then, that the set of messages $M = M_S \times M_P$ where M_S is a set of size $|S|$ (i.e. there are as many messages in M_S as there are gradable properties for each object) and M_P is a set of size two (P for 'polarity'). The players of an Extremity Game will be able to successfully communicate if they can learn to associate each message in M_S with a distinct scale and the two signals in M_P with something akin to *-est* and *least*. As advertised, this setup meets two of the three assumptions in the limitative result (5) – (A2) optimal communication is correct identification of a target object and (A3) messages come from a product space – but drops (A1): because the context varies from play to play of the game, there is no fixed set of objects about which the agents communicate.[20]

4 Experiment

The goal is to show how a simple semantic system like 8 could emerge via a simple dynamics among agents playing an Extremity Game. In particular, we will use *reinforcement learning*:[21] agents make choices, receive some reward (in our case, for successful communication of the target object in context), and adjust their behavior so that they are more likely to make the corresponding choices in the future.

While most approaches to reinforcement learning in signaling games use a variant of a simple algorithm called Roth-Erev learning,[22] such an algorithm will not suffice for present purposes. On this approach, choices are reinforced entirely independently of one another. Two factors of the present setup require a stronger method. On the practical side, there is a combinatorial explosion that comes from having variable contexts with multiple objects that have multiple gradable properties: there are so many contexts that most of them will not be seen often enough for such an algorithm to be effective. On the conceptual side, if

[20] While the agents in an intuitive sense communicate 'about' a fixed set of objects – all objects with $|S|$ gradable properties – each communicative exchange concerns a different subset thereof.

[21] Sutton and Barto (2018)

[22] Roth and Erev (1995)

choices are reinforced entirely independently, there will be no pressure for signals to emerge that group objects based on the degrees of various properties and their relative position on scales in context.

To overcome this limitation, I will use a type of agent with a built-in capacity for stimulus generalization: artificial neural networks.[23] This choice was made because such networks provide a simple, widely used, and somewhat biologically plausible model that has the capacity to generalize. Other approaches to stimulus generalization in learning in signaling games use a method called *spill-over*.[24] In that framework, not only are the actual choices reinforced, but so too are *similar* choices in similar choice points. Exactly how reinforcement works thus depends on definitions of similarity between choices and between states. While some domains provide natural such definitions,[25] it is not immediately obvious how to define how similar one context-target pair is to another in an Extremity Game. Neural networks will learn to treat certain pairs as similar and others not, without the theorist having to hard-wire a definition of similarity into the learning model.[26]

4.1 Methods

A trial of our experiment will consist of some number of iterations of playing an Extremity Game as in (7). The sender and receiver are each neural networks, schematically depicted in Figure 1. They are trained using the REINFORCE algorithm, the simplest in a family of methods known as policy gradient methods.[27] The intuition behind this algorithm is just as before. Consider the sender. The sender is a policy that takes as input a context and a target and outputs a probability distribution over messages (in this case, two distributions: one over M_S and one over M_P). The sender's policy is parameterized by the weights and biases that connect the neurons in the network. Thanks to what is known as the policy gradient theorem, modern variants of stochastic gradient descent can be used to adjust the weights and biases in a way that is guaranteed to make positively reinforced actions sampled from the policy more likely in the future.

We varied the number of dimensions (i.e. gradable properties) between 1 and 3, and ran 10 trials for each. We trained for five-, twenty-, and fifty-thousand mini-batches respectively, where each mini-batch was size 64. In other words, the agents play 64 games in between each update of their policies; this reduces the variance in learning. We also experimented with two different neural architectures for the receiver – called Basic and Attentional – for reasons that will become clear in what follows. We recorded the rolling accuracy over 10 training steps, as well as the accuracy and detailed properties about contexts and signals used on 5000 new games at the end of training.

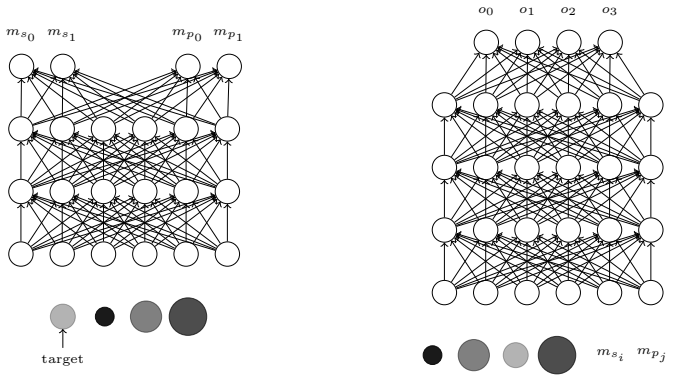
[23] Nielsen (2015) and Goodfellow, Bengio, and Courville (2016)

[24] See O'Connor (2014). The name 'spill-over' comes from Franke (2016).

[25] For instance, if the goal is to choose a point on a line, the distance between the true point and the guessed point is very natural.

[26] See Lazaridou, Peysakhovich, and Baroni (2017) for a similar approach, which inspired the present one. Their contexts consist of two natural images, one of which is the target. The sender chooses one signal from a fixed-sized vocabulary to send to the receiver. While they are interested in whether natural concepts emerge in such a setting, I am focused on less natural input but more complex communication structures in order to explore the emergence of functional vocabulary.

[27] Williams (1992). See chapter 13 of Sutton and Barto (2018) for a modern introduction.



(a) Basic Sender (b) Basic Receiver

Figure 1: Schematic depictions of basic network architectures. The input is on the bottom, followed by a sequence of hidden layers to output layers on the top. The output neurons produce probabilities of choosing the action written above them.

Complete details of the network architectures and training set-up are included in an Appendix. The code and data can be found at https://github.com/shanest/function-words-context.

4.2 Results: Basic Receiver

The learning curves over training for each trial of each dimension, with the Basic Receiver, are plotted in Figure 2. As can be seen, in the one- and two-property cases, the agents learn to communicate nearly perfectly in a relatively short amount of training steps. By contrast, in the three-dimension case, the agents do not regularly achieve a high degree of communicative success after 50000 mini-batches. The mean success rates on 5000 new games at the end of training time are reported in Table 1.

In the one-dimensional case, the context consists of two objects that have a property to different degrees. The successful communication protocol that the agents learn to use reliably sends one signal when the target has the lower degree and the other signal when the target has the higher degree.

dims	mean	std
1	0.975	0.006
2	0.985	0.003
3	0.731	0.062

Table 1: Accuracies on novel games.

In the two-dimensional case, things are not quite as aligned with expectations. Figure 3 shows a typical communication protocol that emerges in the two dimensional case. The colored bars correspond to the particular signals sent. The left column corresponds to M_S and the right column to M_PThe colored bars correspond to the particular signals sent. The left column corresponds to M_S and the right column to M_P. In the top row, the x-axis corresponds to the 'true' dimension of the target object (i.e. the dimension for which the target had an extreme value in context). In the bottom row, the x-axis corresponds to the 'true' polarity of the target object (i.e. whether it had the true property to the least or highest degree).

The bottom-left cell shows an interesting pattern: the message from M_S sent always

Figure 2: Learning curves for basic sender and receiver.

corresponds to the true polarity (minimum or maximum). This is because one message is always sent when the true polarity is 0 (minimum) and the other when the polarity is 1 (maximum). Unfortunately, that the top row shows no such separation implies that no signal is being used to communicate the 'true' dimension. The equal heights of all the bars in the top row imply that the two messages in M_S (left column) and in M_P (right column) are used an equal number of times when the true dimension is 1 and when the true dimension is 0.

In fact, closer inspection reveals the following: the learned communication systems are always 'maximally' separating in the following sense: for any two contexts c, c' and targets o, o', if $o = \arg\min_c s_d(o)$ and $o' = \arg\max_{c'} s_d(o)$ for the same dimension d, then the sender's message for o in c differs from its message for o' in c' in both syntactic positions. This holds true for the 3-dimensional case as well. Figure 4 shows an example learned system. The bottom-right cell shows that the agents do use M_P to distinguish the true direction of the target. But the top-left cell shows that the agents do not associate different signals in M_S with different dimensions: rather, they separate targets in the way just described.

These results show that basic senders and receivers do not, under the REINFORCE algorithm, learn to communicate in accord with the toy semantics in (8). One might think that one of the messages still looks like a superlative morpheme, since it reliably correlates

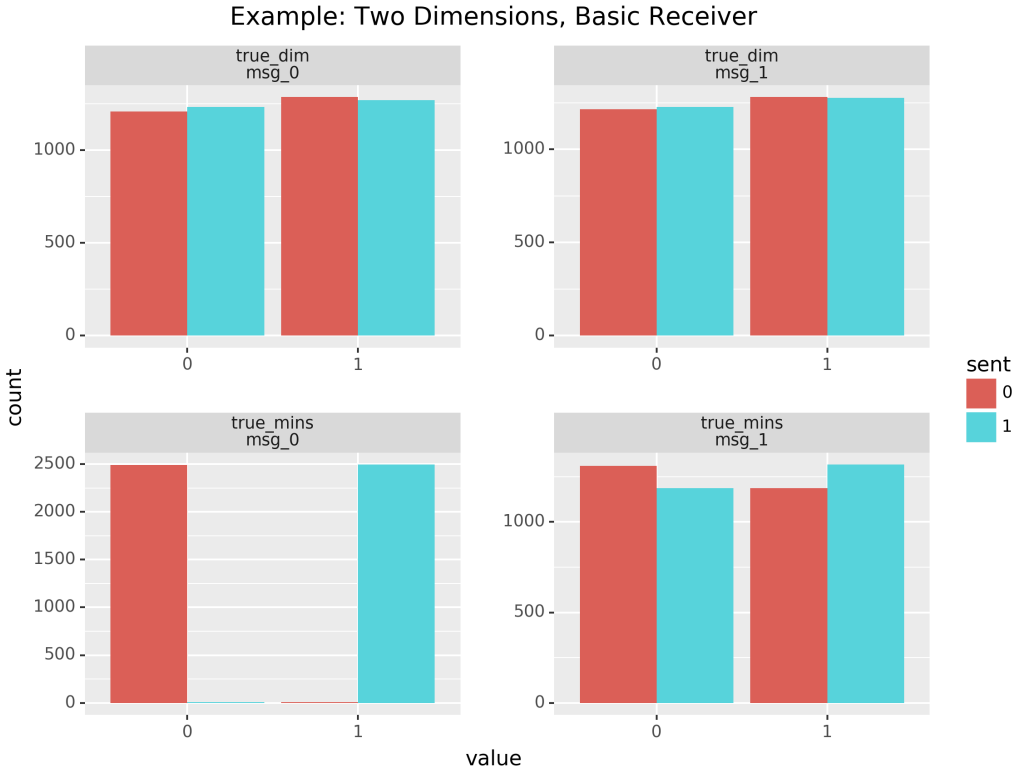


Figure 3: Example communication system with basic receiver and two dimensions.

with the true direction of the target object. While this is indeed very interesting and does show that the networks are clustering objects on the basis of their direction (for example, they never separate on dimension and group together based on direction), given that they do not use the other signal to communicate the true dimension, it does not look like there's non-trivial modification of one linguistic item by another.

4.3 Results: Attentional Receiver

Intuitively, the networks are not learning to use a signal to group objects together based on dimension. This could be for roughly the following reason: in expectation, target objects that differ only in whether they are the minimum/maximum in context on a dimension will actually be farther from each other in Euclidean space than from other objects. Because of this, it could be that the agents use maximally different signals for the two types of target objects.

To help the agents learn to communicate based on the dimension, I will use what is known as an *attention mechanism* in machine learning.[28] Intuitively, a neural network can

[28]See, for instance, Mnih et al. (2014) and Xu et al. (2015).

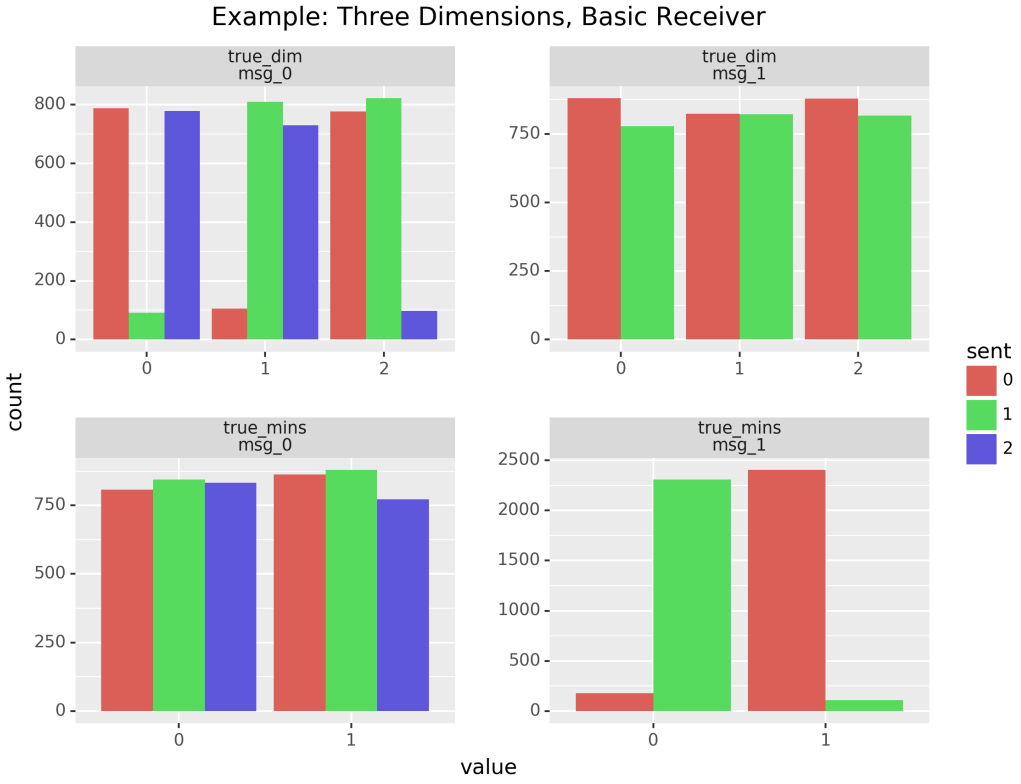


Figure 4: Example communication system with basic receiver and three dimensions.

learn to pay more or less attention to different portions of its input. The network (or a sub-component thereof) computes a weighting of the input positions which is then used to filter the actual input. The weight can be 'hard' – selecting a sub-region of the input – or 'soft' – re-weighting the input so that different nodes are more or less attended to than in the raw input.

One can think of attention as reflecting something like perceptual salience: the network can learn to focus its attention on salient features of its input, since those features are likely to help it solve its task. For instance, a neural image caption generator with attention will likely focus its attention on well-defined objects in an input image. These salient objects are likely to help it generate a plausible caption.

Attentional Receivers, as I will develop them, implement a hard attention mechanism in the following sense. First, they receive as input the context c and the message m_{s_i} from M_S chosen by the sender. On this basis, the receiver *chooses a dimension to attend to*: the input is filtered so that the agent only sees the objects according to one dimension (e.g. size or lightness). Then, the agent uses this attended-to dimension and the message from M_P chosen by the sender to choose a target object. This attention mechanism reflects the perceptual salience of the gradable properties of the objects: it is very natural, for instance,

in the contexts in Section 3, to attend only to the size or the shade of the circles. Figure 5 depicts this architecture.

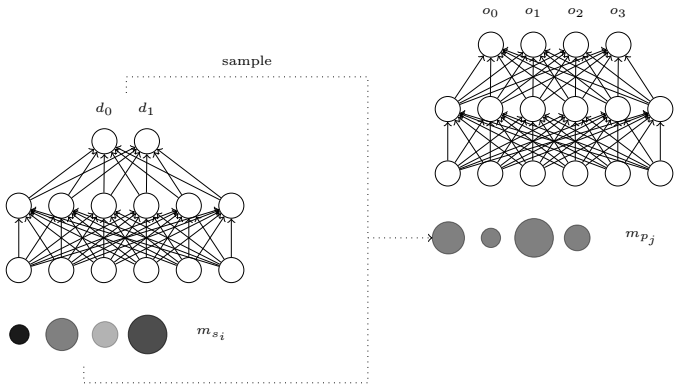


Figure 5: Attentional Receiver architecture, schematically. The receiver first chooses a dimension to attend to, then chooses a target based only on that dimension. In this schematic, the chosen dimension is size; differences in shading have been washed out by the attention mechanism.

The learning curves over training for each trial of each dimension – but with a Basic Sender and Attentional Receiver – are plotted in Figure 6. The mean success rates on 5000 new games at the end of training time are reported in Table 2. As before, in the one- and two-property cases, the agents learn to communicate nearly perfectly in a relatively short amount of training steps. In all cases, it appears that learning is a bit slower than with basic receivers. This makes perfect sense: an attentional receiver has to learn two types of choices to make, as opposed to just one. In the three dimensional case, the attentional receiver achieves a high-degree of accuracy more frequently than the basic receiver, but also gets stuck in sub-optimal states more frequently.

The resulting communication protocols behave exactly like the toy semantics in (8). Figure 7 shows an example protocol in two dimensions. Here, the top-left cell shows that the choice of signal from M_S reliably communicates the true dimension: when the dimension is 0, the sender chooses m_{s_0} and when the dimension is 1, the sender chooses m_{s_1}. Similarly, the bottom-right cell shows that the choice of signal from M_P signal reliably communicates the true direction (i.e. whether the target has the relevant property to the largest or smallest degree). Figure 8 shows an example learned communication system in three dimensions. Again, in complex signals, one signal communicates a dimension, and the other communicates whether the target has the most or least degree on the corresponding scale.

dims	mean	std
1	0.959	0.005
2	0.964	0.005
3	0.697	0.144

Table 2: Accuracies on novel games.

When the agents are communicating in this way, the signals that communicate direction can be interpreted as function words. The signals in M_S reliably communicate a bit of 'content': a dimension. The signals in M_P reliably signal whether the target has the greatest/lowest degree *along that dimension* of all the objects in the context. This is non-trivial

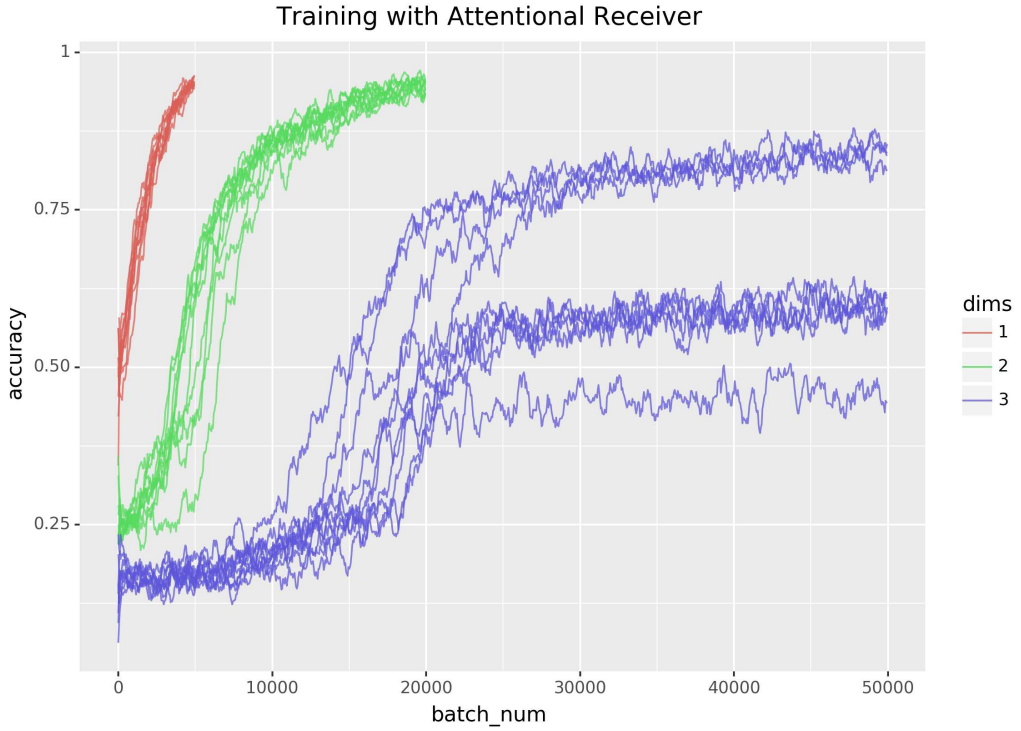


Figure 6: Learning curves for basic sender and attentional receiver.

modification of one linguistic item by another. Thus, when the receiver knows to use one of the signals to attend to a particular dimension in context, the two agents can learn to use their signals in a non-trivially compositional way.

5 Conclusion

Let us take stock. After introducing the distinction between functional and lexical categories, I argued that there are in principle reasons why many extant models of the evolution of compositionality cannot explain the emergence of function words: given their assumptions, they can only explain trivial composition; but non-trivial composition is a necessary precondition for the presence of function words. I then introduced a signaling game with variable contexts consisting of multiple objects with varying gradable properties. Simple reinforcement learning by neural networks – in particular with the ability to pay attention to certain perceptually salient aspects of the input – in this game can generate expressions that are appropriately characterized as function and as content words.

Much work remains to be done. One would like neural architectures that make fewer assumptions about what aspects of the input the receiver pays attention to. A first step in this direction will be to use a soft, as opposed to hard, attention mechanism. A more

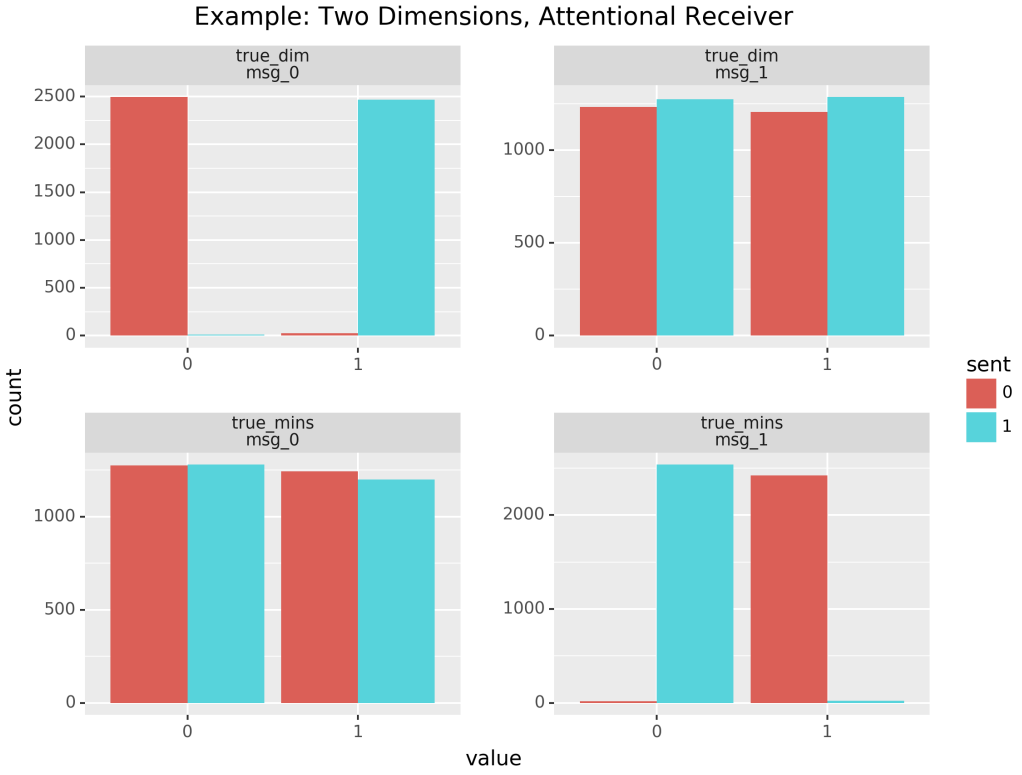


Figure 7: Example communication system with attentional receiver and two dimensions.

thorough hyper-parameter search may also generate more reliable learning results in the higher-dimensional setting. One can also generalize the input so that the networks also have to discover *which dimensions* are relevant for being able to successfully refer to objects across contexts, instead of having it built into the current definition of context. More generally, one would like communication systems like those exhibited here to emerge in the very general setting of communicating by a sequence of symbols with costs for things like vocabulary size and length of messages. All of these exciting avenues remain to be pursued in future work.

References

Barrett, Jeffrey A (2007). "Dynamic Partitioning and the Conventionality of Kinds". In: *Philosophy of Science* 74, pp. 527–546.

— (2009). "The Evolution of Coding in Signaling Games". In: *Theory and Decision* 67.2, pp. 223–237. DOI: 10.1007/s11238-007-9064-0.

Barrett, Jeffrey A, Brian Skyrms, and Calvin Cochran (2018). "Hierarchical Models for the Evolution of Compositional Language". In: *26th Philosophy of Science Association Biennnial Meeting*.

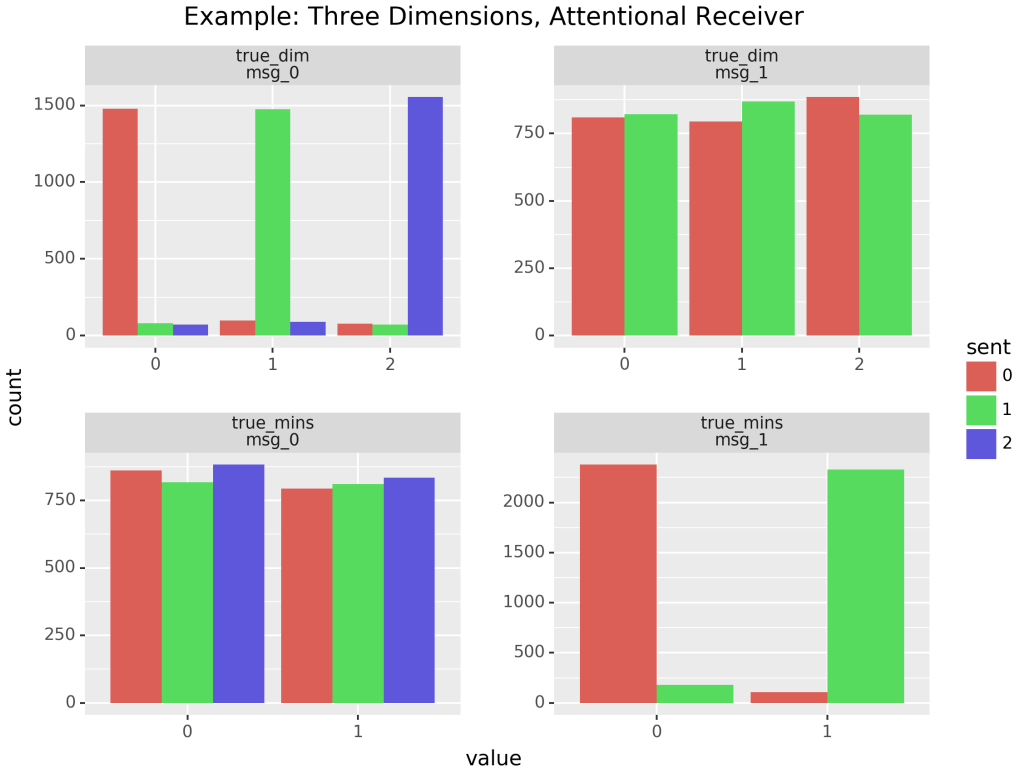


Figure 8: Example communication system with attentional receiver and three dimensions.

Berthet, Mélissa et al. (2018). "Titi monkey alarm sequences: when combining creates meaning". In: *26th Philosophy of Science Association Biennnial Meeting*.

Carnie, Andrew (2006). *Syntax: A Generative Introduction*. Second. Oxford: Blackwell Publishing.

Carroll, Lewis (1871). *Through the Looking-Glass, and What Alice Found There*. Macmillan.

Cäsar, Cristiane et al. (2013). "Titi monkey call sequences vary with predator location and type". In: *Biology Letters* 9.20130535, pp. 2–5. DOI: 10.1098/rsbl.2013.0535.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2016). "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". In: *International Conference of Learning Representations*. URL: http://arxiv.org/abs/1511.07289.

Franke, Michael (2016). "The Evolution of Compositionality in Signaling Games". In: *Journal of Logic, Language and Information*. DOI: 10.1007/s10849-015-9232-5.

Frege, Gottlob (1923). "Logische Untersuchungen. Dritter Teil: Gedankengefüge ("Compound Thoughts")". In: *Beiträge zur Philosophie des deutschen Idealismus* III, pp. 36–51. DOI: 10.1093/mind/LI.202.200.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. The MIT Press. URL: https://www.deeplearningbook.org/.

Graff, Delia (2000). "Shifting Sands: An Interest-Relative Theory of Vagueness". In: *Philosophical Topics* 28.1, pp. 45–81.

Heim, Irene and Angelika Kratzer (1998). *Semantics in Generative Grammar*. Blackwell Textbooks in Linguistics. Wiley-Blackwell.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

Jacobson, Pauline (2014). *Compositional Semantics: An Introduction to the Syntax/Semantics Interface*. Oxford Textbooks in Linguistics. Oxford Univeristy Press.

Janssen, Theo M V (1997). "Compositionality". In: *Handbook of Logic and Language*. Ed. by Johan van Benthem and Alice ter Meulen. Elsevier Science. Chap. 7, pp. 417–473. DOI: 10.1016/B978-044481714-3/50011-4.

Kaplan, David (1978). "Dthat". In: *Syntax and Semantics*. Ed. by Peter Cole. Vol. 9. New York: Academic Press, pp. 212–233.

Kennedy, Christopher (2007). "Vagueness and grammar: the semantics of relative and absolute gradable adjectives". In: *Linguistics and Philosophy* 30, pp. 1–45. DOI: 10.1007/s10988-006-9008-0.

Kennedy, Christopher and Louise McNally (2005). "Scale Structure, Degree Modification, and the Semantics of Gradable Predicates". In: *Language* 81.2, pp. 345–381.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference of Learning Representations (ICLR)*. arXiv: 1412.6980. URL: https://arxiv.org/abs/1412.6980.

Lazaridou, Angeliki, Alexander Peysakhovich, and Marco Baroni (2017). "Multi-Agent Cooperation and the Emergence of (Natural) Language". In: *International Conference of Learning Representations (ICLR2017)*. arXiv: 1612.07182. URL: http://arxiv.org/abs/1612.07182.

Levin, Beth and Malka Rappaport Hovav (2005). *Argument Realization*. Cambridge University Press.

Lewis, David (1969). *Convention*. Blackwell.

— (1988). "Relevant Implication". In: *Theoria* 54.3, pp. 161–174. DOI: 10.1111/j.1755-2567.1988.tb00716.x.

Mnih, Volodymyr et al. (2014). "Recurrent Models of Visual Attention". In: pp. 1–12. arXiv: 1406.6247. URL: http://arxiv.org/abs/1406.6247.

Mordatch, Igor and Pieter Abbeel (2018). "Emergence of Grounded Compositional Language in Multi-Agent Populations". In: *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*. URL: http://arxiv.org/abs/1703.04908.

Muysken, Pieter (2008). *Functional Categories*. Cambridge: Cambridge University Press.

Nielsen, Michael A (2015). *Neural Networks and Deep Learning*. Determination Press. URL: http://neuralnetworksanddeeplearning.com/.

Nowak, Martin A and David C Krakauer (1999). "The evolution of language". In: *Proceedings of the National Academy of Sciences* 96, pp. 8028–8033.

O'Connor, Cailin (2014). "Evolving Perceptual Categories". In: *Philosophy of Science* 81.5, pp. 840–851.

Ouattara, Karim, Alban Lemasson, and Klaus Zuberbühler (2009). "Campbell's monkeys concatenate vocalizations into context-specific call sequences." In: *Proceedings of the National Academy of Sciences* 106.51, pp. 22026–22031. DOI: 10.1073/pnas.0908118106.

Pagin, Peter and Dag Westerståhl (2010a). "Compositionality I: Definitions and Variants." In: *Philosophy Compass* 5.3, pp. 250–264. DOI: 10.1111/j.1747-9991.2009.00228.x.

— (2010b). "Compositionality II: Arguments and Problems." In: *Philosophy Compass* 5.3, pp. 265–282. DOI: 10.1111/j.1747-9991.2009.00229.x.

Partee, Barbara Hall (1995). "Lexical Semantics and Compositionality". In: *Invitation to Cognitive Science, Part 1: Language*. Ed. by Lila Gleitman and Mark Liberman. Cambridge: MIT Press. Chap. 11, pp. 311–360.

Rizzi, Luigi and Guglielmo Cinque (2016). "Functional Categories and Syntactic Theory". In: *Annual Review of Linguistics* 2.1, pp. 139–163. DOI: 10.1146/annurev-linguistics-011415-040827.

Roth, Alvin E. and Ido Erev (1995). "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term". In: *Games and Economic Behavior* 8, pp. 164–212.

Schlenker, Philippe, Emmanuel Chemla, Kate Arnold, et al. (2014). "Monkey semantics: two 'dialects' of Campbell's monkey alarm calls". In: *Linguistics and Philosophy* 37, pp. 439–501. DOI: 10.1007/s10988-014-9155-7.

Schlenker, Philippe, Emmanuel Chemla, Anne M Schel, et al. (2016a). "Formal monkey linguistics". In: *Theoretical Linguistics* 42.1-2, pp. 1–90. DOI: 10.1515/tl-2016-0001.

Schlenker, Philippe, Emmanuel Chemla, Anne M Schel, et al. (2016b). "Formal monkey linguistics: The debate". In: *Theoretical Linguistics* 42.1-2, pp. 173–201. DOI: 10.1515/tl-2016-0010.

Skyrms, Brian (2010). *Signals: Evolution, Learning, and Information*. Oxford University Press.

Steinert-Threlkeld, Shane (2014). "Learning to Use Function Words in Signaling Games". In: *Proceedings of Information Dynamics in Artificial Societies (IDAS-14)*. Ed. by Emiliano Lorini and Laurent Perrussel.

— (2016a). "Compositional Signaling in a Complex World". In: *Journal of Logic, Language and Information* 25.3, pp. 379–397. DOI: 10.1007/s10849-016-9236-9.

— (2016b). "Compositionality and competition in monkey alert calls". In: *Theoretical Linguistics* 42.1-2, pp. 159–171. DOI: 10.1515/tl-2016-0009.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: an introduction*. Second Edi. The MIT Press.

Syrett, K., C. Kennedy, and J. Lidz (2010). "Meaning and Context in Children's Understanding of Gradable Adjectives". In: *Journal of Semantics* 27.1, pp. 1–35. DOI: 10.1093/jos/ffp011.

Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine Learning* 8.3-4, pp. 229–256.

Xu, Kelvin et al. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *International Conference on Machine Learning (ICML 32)*. Ed. by Francis Bach and David Blei, pp. 2048–2057. arXiv: 1502.03044. URL: https://arxiv.org/abs/1502.03044.

Zuberbühler, Klaus (2018). "Combinatorial capacities in primates". In: *Current Opinion in Behavioral Sciences* 21, pp. 161–169. DOI: 10.1016/j.cobeha.2018.03.015.

A Full Experiment Details

For each number of dimensions n, a context has $2n$ objects. Each object is specified by n real numbers, chosen uniformly at random from the interval $(0, 2)$ at steps of 0.1. The values are uniformly subtracted by 1 to center them around 0.

The sender thus has $2n^2$ input nodes. As a convention, the first object for the sender is always the target. It has two hidden layers of 64 nodes each, with exponential linear activation.[29] The final hidden layer is then passed through two linear layers, with output sizes $|M_S|$ and 2, respectively. These are batch normalized[30] and fed into a softmax, to generate distributions over M_S and M_P.

The Basic Receiver receives the context, but with the objects in a random order compared to the sender, and two signals sampled from the sender's output distributions, encoded as one-hot vectors. It then has three rectified linear hidden layers of 64, 64, and 32 units respectively. Then a final linear layer with $2n$ output nodes (one for each target object) is passed through batch normalization and softmax to generate a distribution.

The Attentional Receiver passes the context and a messaged from M_S sampled from the sender through one exponential linear layer of 64 units, before batch normalization and softmax of size n, one for each dimension. A sample is taken from this distribution. The corresponding scalar values for each object along the dimension, together with a message sampled from the sender's distribution over M_P are passed through exponential linear layers of size 64 and 32, before batch normalization and softmax produce a distribution over target objects.

We trained using the REINFORCE algorithm, with mini-batches of size 64, and the Adam optimizer[31] with learning rate $5 \cdot 10^{-4}$. For $n = 1, 2, 3$ dimensions, and each type of receiver, we ran 10 trials of 5000, 20000, and 50000 mini-batches of training. After training, the trained networks then played 5000 versions of the game; the signals chosen, the target chosen, whether it was correct, and what the 'true' dimension and direction (min/max) for identifying the target in context were recorded.

Everything was implemented in PyTorch. The code and data are available at https://github.com/shanest/function-words-context.

[29] Clevert, Unterthiner, and Hochreiter (2016)
[30] Ioffe and Szegedy (2015)
[31] Kingma and Ba (2015)

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association

Seattle, WA; 1-4 November 2018

Version: 31 October 2018

PhilSci ARCHIVE

Table of Contents

Mixed-Effects Modeling and Non-Reductive Explanation

(4975 words)

Abstract: This essay considers a mixed-effects modeling practice and its implications for the philosophical debate surrounding reductive explanation. Mixed-effects modeling is a species of the multilevel modeling practice, where a single model incorporates simultaneously two (or even more) levels of explanatory variables to explain a phenomenon of interest. I argue that this practice makes the position of explanatory reductionism held by many philosophers untenable, because it violates two central tenets of explanatory reductionism: single level preference and lower-level obsession.

1. Introduction

Explanatory reductionism is the position which holds that, given a relatively higher-level phenomenon (or state, event, process, etc.), it can be reductively explained by a relatively lower-level feature (Kaiser 2015, 97; see also Sarkar 1998; Weber 2005; Rosenberg 2006; Waters 2008).[1] Though philosophers tend to have slightly different conceptions of the position, two central tenets of the position can still be extracted:[2]

> **Single level preference:** a phenomenon of interest can be fully explained by invoking features that reside at a single, well-defined level of analysis (e.g., molecular level in biology).

[1] According to Sarkar (1998), explanatory reduction is an epistemological thesis which is distinguished from constitutive (ontological) and theory reductionism theses. Kaiser further distinguishes two sub-types of explanatory reduction: (a) "a relation between a higher-level explanation and a lower-level explanation of the same phenomenon" (2015, 97); (b) individual explanations, i.e., given a relatively higher-level phenomenon, it can be reductively explained by a relatively lower-level feature (*Ibid*., 97). This essay will focus on the second sub-type. Besides, when referring to levels I mean either hierarchical organization such as universities, faculties, departments etc., or functional organization such as organs, tissues, cells etc. When referring to scales I mean spatial or temporal scaling where levels are not so clearly delimited.

[2] Similar summary of the position can be found in Sober (1999).

> **Lower-level obsession:** lower-level features always provide the most significant and detailed explanation of the phenomenon in question, so a lower-level explanation is always better than a higher-level explanation.

Philosophers sometimes express these two tenets explicitly in their work. For example, Alex Rosenberg holds that "[…] there is a full and complete explanation of every biological fact, state, event, process, trend, or generalization, and that this explanation will cite only the interaction of macromolecules to provide this explanation" (Rosenberg 2006, 12). Marcel Weber expresses a similar idea in his explanatory hegemony thesis, according to which it's always some lower-level physicochemical laws (or principles) that ultimately do the explanatory work in experimental biology (Weber 2005, 18-50). John Bickle attempts to motivate a 'ruthless' reduction of psychological phenomena (e.g., memory) to the molecular level (Bickle 2003).

However, many philosophers have questioned the plausibility of the position on the basis of scientific practice (Hull 1972; Craver 2007; Bechtel 2010; Brigandt 2010; Hüttemann and Love 2011; Kaiser 2015). To counter that position, some authors have pointed to the relevance of an important practice that has not received sufficient attention before: multiscale or multilevel modeling or sometimes called integrative modeling approach, where a set of distinct models ranging over multiple levels or scales—including the macro-phenomenon level/scale—are involved in explaining a (often complex) phenomenon of interest

(Mitchell 2003, 2009; Craver 2007; Brigandt 2010, 2013a, 2013b; Knuuttila 2011; Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017). Often these models work together by providing diverse constraints on the potential space of representation (Knuuttila and Loettgers 2010; Knuuttila 2011; Green 2013).

This multilevel modeling surely casts some doubt on explanatory reductionism, for it seems unclear what reductively explains what—all those facts in the set of models ranging over different levels/scales are involved in doing some explanatory work. However, there is a species of multilevel modeling that has slipped away from most philosophers' sights: mixed-effects modeling (MEM hereafter)—also called multilevel regression modeling, hierarchical linear modeling, etc.—in which a single model incorporating simultaneously two (or even more) levels of variables is used to explain a phenomenon. For a mixed-effects model to explain, features of the so-called reducing and reduced levels must be simultaneously incorporated into the model, that is, they must go hand in hand.

MEM deserves special attention because it sheds new light on the reductionism-antireductionism debate by showing that (a) a mixed-effects model violating the two central tenets of explanatory reductionism can provide successful explanation, and (b) a single mixed-effects model without integrating with other epistemic means can also provide such successful explanation. Therefore, MEM first further challenges the explanatory reductionist position, and

second offers a novel perspective bolstering the multilevel/multiscale integrative approach discussed by many philosophers.

The essay proceeds as follows. Section 2 discusses the challenges faced by the traditional single-level modeling approach, and examines the reasons why the MEM approach is preferable in dealing with these challenges. Section 3 describes a MEM practice using a concrete model. Section 4 elaborates on the implications of MEM for the explanatory reductionism debate. Finally, Section 5 considers potential objections to my viewpoint.

2. Challenges to Reductive Explanatory Strategies

In many fields (e.g., biological, social and behavioral sciences) scientists find that the data collected show an intrinsically hierarchical or nested feature. Consider a simple example: we might be interested in examining relationships between students' achievement at school (A hereafter) and the time they invest in studying (T).[3] In conducting such a research, we might collect data from different classes (say 5 classes in total), with each class providing the same number of samples (say 10 students in each class). The data collected among classes might be taken for granted to be independent. Then we may use certain traditional statistical techniques such as ordinary least-squares (OLS) to analyze the data and build a linear relationship between A and T.

[3] For scientific studies of this kind, see Schagen (1990), Wang and Hsieh (2012), and Maxwell et al. (2017).

However, this single-level reductive analysis can lead to misleading results, because it ignores the possibility that students within a class may be more similar to each other in important aspects than students from different classes. In other words, each group (class) may have its own features relevant to the relationship between A and T that the other groups lack. Hence, the data collected from the students are in fact not independent, i.e., the subjects are not randomly sampled, because the individuals (students) are clustered within groups (classes). In technical terms, we say our analysis may fall prey to the *atomistic fallacy* where we base our analysis solely on the individual level—i.e., we reduce all the group-level features to the individuals. Therefore, traditional OLS techniques such as multiple regression cannot be employed in this context, because the case under consideration violates a fundamental assumption of these techniques: the independence of observations (Nezlek 2008, 843).

Conversely, we may face the same problem the other way around if we fail to consider the inherently nested nature of the data. Consider the student-achievement-at-school case again. We may observe that in classes where the time of study invested by students is very high, the achievements of the students are also very high. Given such an observation, we may reason that students who invest a lot of time in studying would be more likely to get higher achievements at school. However, this inference commits the *ecological fallacy*, because it attributes the relationship observed at the group-level to the individual-level (Freedman 1999). The individuals may exhibit within-group differences that the single group-level analysis fails to capture. In technical terms, this inference flaws

because it reduces the variability in achievement at the individual-level to a group-level variable, and the subsequent analysis is solely based on group's mean achievement results (Heck and Thomas 2015, 3). Again, traditional statistical techniques such as multiple regression cannot be employed in this context.

In sum, a single-level modeling approach that disrespects the multilevel data structure can commit either an atomistic or an ecological fallacy. Confronted with these problems, one response is to 'tailor' the traditional statistical techniques by, e.g., adding an effect variable to the model which indicates the grouping of the individuals. However, many have argued that this approach is unpromising because it may give rise to enormous new problems (Luke 2004; Nezlek 2008; Heck and Thomas 2015). Alternatively, scientists have developed a new framework that takes the multilevel data structure into full consideration, i.e., the MEM approach, to which we now turn.

3. Case Study: A Mixed-Effects Model

Depending on different conceptual and methodological roots we have two broad categories of MEM approaches: the multilevel regression approach and the structural equation modeling approach. The former usually focuses on direct effects of predictor variables on (typically) a single dependent variable, while the latter usually involves latent variables defined by observed indicators (for details see Heck and Thomas 2015). For the purpose of this essay's arguments, I will concentrate on the first kind.

Consider the student-achievement-at-school example again. Since students are typically clustered in different classes, a student's achievement at school may be both influenced by her own features (e.g., time invested in studying) and her class's features (e.g., size of the class). Hence here comes two levels of analysis: the individual-level (level-1) and the group-level (level-2), and individuals ($i=1,2,\ldots,N$) are clustered in level-2 groups ($j=1,2,\ldots,n$).[4] Now suppose that students' achievements at school are represented as scores they get in the exam. The effect of time invested in studying on scores can be described as follows:

$$Y_{ij}=\beta_{0j}+\beta_{1j}X_{ij}+\varepsilon_{ij} \qquad (1)$$

where Y_{ij} refers to the score of individual i in the jth group, β_{0j} is a level-1 intercept representing the mean of scores for the jth group, β_{1j} a level-1 slope (i.e., different effects of study time on scores) for the predictor variable X_{ij}, and the residual component (i.e., an error term) ε_{ij} the deviation of individual i's score from the level-2 mean in the jth group. Equation (1) looks like a multiple regression model; however, the subscript j reveals that there is a group-level incorporated in the model. It can also be seen from this equation that both the intercept β_{0j} and slope β_{1j} can vary across the level-2 units, that is, different groups can have different intercepts and slopes.

[4] Note that, for instructive purposes, our case involves only two levels; however, the MEM approach can in principle be extended to many more levels.

The most remarkable thing of MEM is that we treat both the intercept and slope at level-1 as dependent variables (i.e., outcomes) of level-2 predictor variables. So here we write the following equations expressing the relationships between the level-1 parameters and level-2 predictors:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad (2)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \quad (3)$$

where β_{0j} refers to the level-1 intercept in level-2 unit j, γ_{00} denotes the mean value of the level-1 intercept, controlling for the level-2 predictor W_j, γ_{01} the slope for the level-2 variable W_j, and u_{0j} the error (i.e., the random variability) for unit j. Also, β_{1j} refers to the level-1 slope in level-2 unit j, γ_{10} the mean value of the level-1 slope controlling for the level-2 predictor W_j, γ_{11} the effect of the level-2 predictor W_j, and u_{1j} the error for unit j.

Equations (2) and (3) have specific meanings and purposes. They express how the level-1 parameters, i.e., intercept or slope, are functions of level-2 predictors and variability. They aim to explain variations in the randomly varying intercepts or slopes by adding one (or more) group-level predictor to the model. These expressions are based on the idea that the group-level characteristics such as group size may impact the strength of the within-group effect of study time on

scores. This kind of effect is called a *cross-level interaction* for it involves the impact of variables at one level of a data hierarchy on relationships at another level. We will discuss this in detail in the next section.

Now we combine equations (1), (2) and (3) by substituting the level-2 parts of the model into the level-1 equation. We finally obtain the following equation:

$$Y_{ij} = \left[\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + \gamma_{11} X_{ij} W_j\right] + \left[u_{1j} X_{ij} + u_{0j} + \varepsilon_{ij}\right] \qquad (4)$$

This equation can be simply understood that Y_{ij} is made up of two components: the fixed-effect part expressed by the first four terms and the random-effect part expressed by the last three terms. Note that the term $\gamma_{11} X_{ij} W_j$ denotes a cross-level interaction between level-1 and level-2 variables, which is defined as the impact of a level-2 variable on the relationship between a level-1 predictor and the outcome Y_{ij}. We have 7 parameters to estimate in (4), they are four fixed effects: intercept, within-group predictor, between-group predictor and cross-level interaction, two random effects: the randomly varying intercept and slope, and a level-1 residual.

Now a mixed-effects model has been built, and the next step is to estimate the parameters of the model. However, we will skip this step and turn to explore the philosophical implications of the modeling practice relevant to the explanatory reductionism debate.

4. Implications for the Explanatory Reductionism Debate

Looking closely into the MEM practice, we find that a couple of important philosophical implications for the explanatory reductionism debate can be drawn.

4.1. All levels are indispensable

The first, and most obvious, feature of MEM is that it routinely involves many levels of analysis in a single model, and all these levels are indispensable to the model in the sense that no level can be reduced to or replaced by the other levels. These levels consist of both the so-called reducing level in the reductionist's terminology, typically a lower-level that attempts to reduce another level, and the reduced level, typically a higher-level to be reduced by the reducing level. In our student-achievement-at-school case, for example, a reductionist may state that the group-level will be regarded as the reduced level whereas the student-level as the reducing level.

The indispensability of each level in the model can be understood in two related ways. First, due to the nested nature of data, only when we incorporate different levels of analyses to the model can we avoid either the atomistic or ecological fallacy discussed in Section 2. As discussed in the student-achievement-at-school example where students are clustered in different classes (in the manner that students from the same class may be more similar to each other in important aspects than students from different classes), reducing all the analyses to the level of individual students can simply miss the important

information associated with group-level features and thus lead to misleading results. Although it's true that the problem might be partially mitigated by tailoring traditional single-level analytical techniques such as multiple regression, it's also true that this somewhat ad hoc maneuver can simply bring about various new vexing and recalcitrant issues (Luke 2004; Nezlek 2008; Heck and Thomas 2015).

Second, the problem can also be viewed from the perspective of identifying explanatory variables. In building a mixed-effects model, the main consideration is often to find a couple of variables that may play the role of explaining the pattern or phenomenon observed in the data. Here a modeler must be clear about how to assign explanatory variables, for instance, she must consider if there are different levels of analyses and, if so, which explanatory variables should be assigned to what levels, and so on. These considerations may come before her model building because of background knowledge, which paves the way for her to develop a conceptual framework for investigating the problem of interest. However, without such a clear and rigorous consideration of identifying and assigning multilevel explanatory variables, an analysis can flaw simply because it confounds variables at different levels.

Respecting the multilevel nature of explanatory variables has another advantage: "Through examining the variation in outcomes that exists at different levels of the data hierarchy, we can develop more refined theories about how explanatory variables at each level contribute to variation in outcomes" (Heck and Thomas 2015, 33). In other words, in respecting the multilevel nature of

explanatory variables, we get a clear idea of how, and to what degrees, explanatory variables at different levels contribute to variation in outcomes. If these variables do contribute to variation in outcomes, as it always happens in MEM, then the situation suggests an image of *explanatory indispensability*: all the explanatory variables at different levels are indispensable to explaining the pattern or phenomenon of interest.

Given these considerations, therefore, one implication for the explanatory reductionism debate becomes clear: it isn't always the case that, given a relatively higher-level phenomenon it can be reductively explained by a relatively lower-level feature. Rather, in cases where the data show a nested structure or, put differently, the phenomenon suggests multilevel explanatory variables, we routinely combine the higher-level with the lower-level in a single (explanatory) model. As a result, one fundamental tenet of explanatory reductionism is violated: single level preference.

4.2. Interactions between levels

Another crucial feature of multilevel modeling is its emphasis on *a cross-level interaction*, which is defined as

> "The potential effects variables at one level of a data hierarchy have on relationships at another level […]. Hence, the presence of a cross-level interaction implies that the magnitude of a relationship observed within

> groups is dependent on contextual or organizational features defined by higher-level units". (Heck and Thomas 2015, 42-43)

Remember that there is a term $\gamma_{11} X_{ij} W_j$ in our mixed-effects model discussed in Section 3, which indicates the cross-level interaction between the group-level and the individual-level. More specifically, this term can be best construed as the impact of a group-level variable, e.g., group size, upon the individual-level relationship between a predictor, e.g., study time, and the outcome, e.g., students' scores.

The cross-level interaction points to the plain fact that an organization or a system can somehow influence its members or components by constraining how they behave within the organization or system. This doesn't necessarily imply top-down causation (Section 5.3 will turn back to this point). Within the context of scientific explanation, however, it does imply that it isn't simply that characteristics at different levels separately contribute to variation in outcomes, but rather that they interact in producing variation in outcomes. In other words, the pattern or phenomenon to be explained can be understood as generated by the interaction between explanatory variables at different levels. Therefore, to properly explain the phenomenon of interest, we need not only have a clear idea of how to assign explanatory variables to different levels but also an unequivocal conception of whether these explanatory variables may interact.

Different models can be built depending on different considerations of the cross-level interaction. To see this, consider the student-achievement-at-school

example again. In some experiment setting we may assume that there was no cross-level interaction between group-level characteristics and the individual-level relationship (between study time and scores). In such a situation, we kept the effect of individual study time on scores the same across different classes, i.e., we kept the slope constant across classes. In the meanwhile, we treated another group-level variable (i.e., intercept) as varying across classes, i.e., different classes have different average scores. So, this is a case where we have a clear idea of how to assign explanatory variables but no consideration of the cross-level interaction. Nonetheless, in a different experiment setting we may assume that there existed cross-level interaction, and hence the effect of individual study time on scores can no longer be kept constant across different classes. At the same time, we treated another group-level variable (i.e., intercept) as varying across classes. Hence, this is a case where we have both a clear idea of how to assign explanatory variables and a consideration of the cross-level interaction. Corresponding to these two different scenarios, two different mixed-effects models can be built, as shown below:

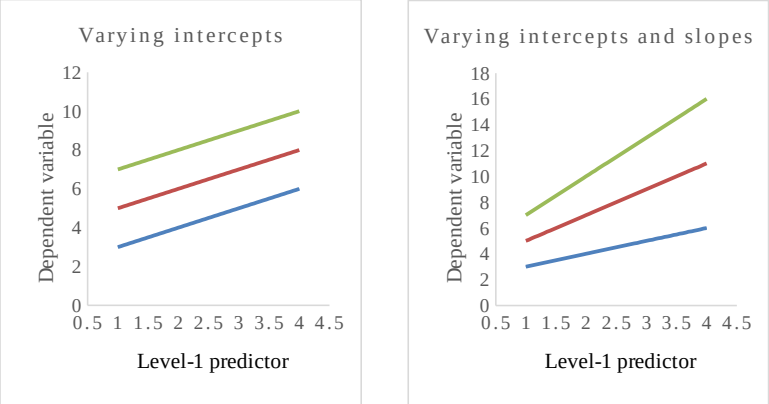


Figure 1. Two different models showing varying intercepts or varying intercepts and slopes, respectively. Three lines represent three classes. This figure is adapted from Luke (2004, 12).

Given such a cross-level interaction, therefore, the explanatory reductionist position has been further challenged. This is because any reductive explanation that privileges one level of analysis—usually the lower-level—over the others falls short of capturing this kind of interaction between levels. If they fail to do so, then they are missing important terms relevant to explaining the phenomenon of interest. As a consequence, a mixed-effects model involving interactions between levels simultaneously violates the two fundamental pillars of explanatory reductionism: first, it violates single level preference because it involves multilevel explanatory variables in explaining phenomena, and second, it violates lower-level obsession because it privileges no levels—all levels are interactively engaged in producing outcomes.

5. Potential Objections

This section considers two potential objections.

5.1. In-principle argument

One argument that resurfaces all the time in the reductionism-versus-antireductionism debate is the in-principle argument, the core of which is that even if reductive explanations in a field of study are not available for the time being, it doesn't follow that we won't obtain them someday (e.g., Sober 1999; Rosenberg 2006). Therefore, according to some reductionists, the gap between current-science and future-science is simply a matter of time, for advancement in techniques, experimentation and data collecting can surely fill in the gap.

However, I think the argument flaws. To begin with, advancement in techniques, experimentation and data collecting isn't always followed by reductive explanations. For example, in our MEM discussed in Section 3, even if the data about the individual-level is available and sufficiently detailed, it isn't the case that we explain the phenomenon of interest in terms of the data from the individual-level alone. Consider another example: in dealing with problems associated with complex systems in systems biology, even though large-scale experimentation (e.g., via computational simulation) can be conducted and high throughput data arranging over multiple scales/levels can be collected, a bottom-up reductive approach must be integrated with a top-down perspective so as to

produce useful explanations or predictions (Green 2013; Green and Batterman 2017; Gross and Green 2017).

Nevertheless, reductionists may reply that the situations presented above only constitute an in-practice impediment, for it doesn't undermine the *possibility* that lower-level reductive explanations, typically provided by some form of 'final science', will be available someday. Let us dwell on the notion of possibility a bit longer. The possibility here may be construed as a *logical possibility* (Green and Batterman 2017, 21; see also Batterman 2017). Nonetheless, if it's merely logically possible that there will be some final science providing only reductive explanations, then nothing can exclude another logical possibility that there will be some 'mixed-science' providing only multilevel explanations. After all, how can we decide which logical possibility is more possible (or logically more possible)? I doubt that logic alone could provide anything useful in justifying which possibility is more possible, and that appealing to logical possibility could offer anything insightful in helping us understand how science proceeds. As Batterman puts, "Appeals to the possibility of *in principle* derivations rarely, if ever, come with even the slightest suggestion about how the derivations are supposed to go" (2017, 12; author's emphasis).

Another interpretation of possibility may be associated with real possibilities, referring to the actual cases of reductive explanations happening in science. Unfortunately, I don't think the real scenario in science speaks for the reductionist under this interpretation. Though it's impossible to calculate the absolute cases of non-reductive explanations occurring in science, a cursive look at scientific

practice can tell that a large portion of scientific explanations proceeds in a non-reductive fashion, as suggested by multilevel modeling (Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017; Mitchell and Gronenborn 2017). Moreover, even in areas such as physics which was regarded as a paradigm for the reductionist stance, progressive explanatory reduction doesn't always happen (Green and Batterman 2017; Batterman 2017).

In sum, we have shown that the in-principle argument fails for it neither offers help in understanding how science proceeds if it's construed as implying a logical possibility, nor goes in tune with scientific practice if it's construed as implying real possibilities.

5.2. Top-down causation

In Section 3 we have shown that there is a cross-level interaction taking the form that higher-level features may impact lower-level features. A worry arises: Does this imply top-down causation?

My answer to this question is twofold. First, it's clear that this short essay isn't aimed to engage in the philosophical debate about whether, and in what sense, there exists top-down causation (see Craver and Bechtel 2007; Kaiser 2015; Bechtel 2017). Second, what we can do now is to show that the cross-level interaction is a clear and well-defined concept in multilevel modeling. It unambiguously means the constraints on the lower-level processes exerted by the higher-level parameters (Green and Batterman 2017). In our multilevel modeling

discussed in Section 3, we have shown that group-level features may impact some individual-level features through the way that each group possesses its own feature relevant to explaining the differences at the individual-level across groups. This idea is incorporated into the mixed-effects model by assigning some explanatory variables to the group-level and a cross-level interaction term to the model.

The idea of cross-level-interaction-as-constraint is widely accepted in multilevel modeling broadly construed, where constraint is usually expressed in the form of initial and/or boundary conditions. For example, in modeling cardiac rhythms, due to "the influences of initial and boundary conditions on the solutions of the differential equations used to represent the lower level process" (Noble 2012, 55; Cf. Green and Batterman 2017, 32), a model cannot simply narrowly focus on the level of proteins and DNA but must also consider the levels of cell and tissue working as constraints. The same story happens in cancer research, where scientists are advocating the idea that tumor development can be better understood if we consider the varying constraints exerted by tissue (Nelson and Bissel 2006; Shawky and Davidson 2015; Cf. Green and Batterman 2017, 32).

6. conclusion

This essay has shown that no-reductive explanations involving many levels predominate in areas where the systems under consideration exhibit a hierarchical structure. These explanations violate the fundamental pillars of explanatory

reductionism: single level preference and lower-level obsession. Traditional single-level reductive approaches fall short of capturing systems of this kind because they face the challenges of committing either the atomistic or ecological fallacy.

References

Batterman, Robert. 2013. The "Tyranny of Scales." In *The Oxford Handbook of Philosophy of Physics*, ed. Robert Batterman, 255-286. Oxford: Oxford University Press.

———. 2017. "Autonomy of Theories: An Explanatory Problem." *Noûs* 1-16.

Bechtel, William. 2010. "The Downs and Ups of Mechanistic Research: Circadian Rhythm Research as an Exemplar." *Erkenntnis* 73:313–328.

———. 2017. "Explicating Top-Down Causation Using Networks and Dynamics." *Philosophy of Science* 84:253–274.

Bickle, John. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Brigandt, Ingo. 2010. "Beyond Reductionism and Pluralism: Toward an Epistemology of Explanatory Integration in Biology." *Erkenntnis* 73 (3): 295-311.

———. 2013a. "Explanation in Biology: Reduction, Pluralism, and Explanatory Aims." *Science and Education* 22:69–91.

———. 2013b. "Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:461–465.

Craver, Carl. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, Carl, and William Bechtel. 2007. "Top-down Causation without Top-Down Causes." *Biology and Philosophy* 22:547–563.

Freedman, David. 1999. "Ecological Inference and the Ecological Fallacy." In *International Encyclopedia of the Social and Behavioral* Sciences, vol. 6, ed. Neil Smelser, and Paul Baltes, 4027–4030. New York: Elsevier.

Green, Sara. 2013. "When One Model Isn't Enough: Combining Epistemic Tools in Systems Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:170–180.

Green, Sara, and Robert Batterman. 2017. "Biology Meets Physics: Reductionism and Multi-Scale Modeling of Morphogenesis." *Studies in History and Philosophy of Biological and Biomedical Sciences* 61:20–34.

Gross, Fridolin, and Sara Green. 2017. "The Sum of the Parts: Large-Scale Modeling in Systems Biology." *Philosophy, Theory, and Practice in Biology* 9: (10).

Heck, Ronald, and Scott Thomas. 2015. *An Introduction to Multilevel Modeling Techniques* (3rd Edition). New York: Routledge.

Hull, David. 1972. "Reductionism in Genetics—Biology or Philosophy?" *Philosophy of Science* 39 (4): 491-499.

Hüttemann, Andreas, and Alan Love. 2011. "Aspects of Reductive Explanation in Biological Science: Intrinsicality, Fundamentality, and Temporality." *British Journal for the Philosophy of Science* 62 (3): 519-549.

Kaiser, Marie. 2015. *Reductive Explanation in the Biological Sciences*. Springer.

Knuuttila, Tarja. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Studies in History and Philosophy of Science Part A* 42:262–271.

Luke, Douglas. 2004. *Multilevel Modeling*. London: SAGE Publications, Inc.

Maxwell, Sophie, Katherine Reynolds, Eunro Lee, et al. 2017. “The Impact of School Climate and School Identification on Academic Achievement: Multilevel Modeling with Student and Teacher Data.” *Frontiers in Psychology* 8:2069.

Mitchell, Sandra. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

———. 2009. *Unsimple Truths: Science, Complexity, and Policy*. Chicago: University of Chicago Press.

Nezlek, John. 2008. “An Introduction to Multilevel Modeling for Social and Personality Psychology.” *Social and Personality Psychology Compass* 2/2 (2008):842–860.

Noble, Daniel. 2012. “A Theory of Biological Relativity: No Privileged Level of Causation.” *Interface Focus* 2(1):55–64.

O’Malley Malley, Ingo Brigandt, Alan Love, et al. 2014. “Multilevel Research Strategies and Biological Systems.” *Philosophy of Science* 81:811–828.

Rosenberg, Alex. 2006. *Darwinian Reductionism, or How to Stop Worrying and Love Molecular Biology*. Chicago: University of Chicago Press.

Sarkar, Sahotra. 1998. *Genetics and Reductionism*. Cambridge: Cambridge University Press.

Schagen, I. P. 1990. “Analysis of the Effects of School Variables Using Multilevel Models.” *Educational Studies* 16:61–73.

Shawky, Joseph, and Lance Davidson. 2015. “Tissue Mechanics and Adhesion during Embryo Development.” *Developmental Biology* 401(1):152–164.

Sober, Elliot. 1999. “The Multiple Realizability Argument against Reductionism.” *Philosophy of science* 66:542–564.

Wang, Yau-De, and Hui-Hsien Hsieh. 2012. “Toward a Better Understanding of the Link Between Ethical Climate and Job Satisfaction: A Multilevel Analysis.” *Journal of Business Ethics* 105:535–545.

Waters, C. Kenneth. 2008. “Beyond Theoretical Reduction and Layer-Cake Antireduction: How DNA Retooled Genetics and Transformed Biological Practice”. In *The Oxford Handbook of Philosophy of Biology*, ed. Michael Ruse, 238-262. New York: Oxford University Press.

Weber, Marcel. 2005. *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

The Universe Never Had a Chance

C. D. McCoy[‡]

1 March 2018

Abstract

Demarest asserts that we have good evidence for the existence and nature of an initial chance event for the universe. I claim that we have no such evidence and no knowledge of its supposed nature. Against relevant comparison classes her initial chance account is no better, and in some ways worse, than its alternatives.
Word Count: 4712

1 Introduction

Although cosmology, the study of the universe's evolution, has largely become a province of physics, philosophical speculation concerning cosmogony, the study of the origin of the universe, continues up to the present. Certainly, many believe that science has settled this too by way of the well-known and well-confirmed big bang model of the universe. According to the big bang account the universe began in a extremely hot, dense state, composed of all the different manifestations of energy that we know. Indeed, time itself began with the big bang. Yet, properly speaking, the universe's past singularity is not some event in spacetime according to the general theory of relativity. In cosmological models this hot dense state called the big bang is generally understood instead as just a very early stage of the universe's evolution, i.e. properly a part of cosmology and not cosmogony. While we may be highly confident that the entire big bang story is correct back to a very early time, our confidence should at some point decrease as we near the supposed "first moment". Thus there remains world enough and time to engage in traditional philosophical and scientific speculations about cosmogony and cosmology alike. Were there previous stages to the universe? What brought the universe into existence? What was the character of this initial happening (should it in fact exist)?

The ubiquity of probabilities in modern physical theories, e.g. quantum mechanics and statistical mechanics, has led some to wonder as well how chance should fit into our

[*]**Acknowledgements:** Pending.

[†]School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK. email: casey.mccoy@ed.ac.uk

cosmogonical worldview. In this vein, Demarest (2016) argues that the probabilities of all events in a(n ostensibly) deterministic universe can be derived from an initial chance event and, what's more, that "we have good evidence of its existence and nature." In this paper I aim to dispute these latter claims. I argue that we do not have any evidence at all of an initial chance event in a big bang universe as described above, much less of its nature. What we rather have in Demarest's account is just a particular way of interpreting probabilistic theories, where all probabilities are taken to derive from ontic chances pertaining to the particular genesis of the relevant physical system, e.g. the universe as a whole. I claim that this interpretation, while coherent, should be disfavored in cosmology—we should rather say that *the universe never had a chance*.[1] Along the way I will make several clarifying remarks concerning the relation of chance and determinism, cosmological probabilities, and alternative interpretations of statistical and quantum mechanics.

2 Chance and Determinism in Physical Theory

By the *world* metaphysicians usually mean something like "the maximally inclusive entity whose parts are all the things that exist." Of course terminology varies. This particular rendering comes from Schaffer (2010, 33), who instead chooses to call this entity the *cosmos*. Cosmologists do not usually call their object of study the cosmos; more commonly they say that they study the *universe*. In *Cosmology: The Science of the Universe*, Harrison explicitly notes the philosophical and historical dimensions of the world taken in its broadest sense, designating this world as a whole the *Universe*. Cosmology, according to Harrison, is the study of universes, by which he means particular models of the Universe (Harrison, 2000, Ch. 1). Cosmological models are the particular concern of physical cosmologists; they are physical models of the Universe, which describe especially its large-scale structure and the evolution thereof.

In what follows I employ these terminologies in the following way. By the *world* I designate the locus of (principally) metaphysical questions concerning the Universe. Is the world deterministic? Is it chancy? By the *universe* I designate the locus of principally physical questions concerning the Universe. How did the big bang universe begin? How will it end? These are questions to which the big bang model should provide an answer.

I do not mean, of course, to introduce an admittedly arbitrary distinction between science and metaphysics by differentiating universes and worlds. Indeed, when one asks whether the world is deterministic, many metaphysicians of science would look first to models of the Universe to help decide the question. Wüthrich for example remarks, matter-of-factly, that "this metaphysical question deflates into the question of whether our best physical *theories* entail that the world is deterministic or indeterministic" (Wüthrich, 2011, 366).

[1] There are several senses, in fact, in which this claim is true. Cosmology suggests that the inevitable fate of the universe is to become ever more sparse and empty through the accelerated expansion of space under the influence of dark energy.

Indeed, many discussions of determinism adopt the approach mentioned by Wüthrich. Let *determinism* denote the thesis that the world is deterministic. Then, following for example (Lewis, 1983, 360), a world is *deterministic* if and only if the laws of that world are deterministic. To determine whether the laws of the universe are deterministic, we must look to our theories of which those laws are part and ask whether those laws taken together should be considered deterministic. It is by no means a straightforward matter to decide whether a given physical theory is deterministic of course. Even the classic example of deterministic physics, Newtonian mechanics, admits many counterexamples against its putative determinism (Earman, 1986; Norton, 2008). General relativity as well seemingly permits indeterministic phenomena in the form of causal pathologies (closed timelike curves) (Earman, 1995) and, if the hole argument is to be believed, is hopelessly rife with indeterminism (Earman and Norton, 1987).

Although classical theories like classical mechanics and general relativity are nevertheless debatably deterministic, surely probabilistic theories like quantum mechanics are properly characterized as indeterministic (at least so long as the probabilities involved are objective features of the world). Yet various interpretations of probabilistic theories seek to avoid indeterminism even here, where it seems unassailable, by characterizing probabilities as merely epistemic or subjective, or else by presenting them as fully deterministic theories (as in the Bohmian interpretation of quantum mechanics). Philosophers have raised serious concerns, however, over how one can truly understand probabilities in deterministic theories, an issue that has been termed the "paradox of deterministic probabilities" (Loewer, 2001; Winsberg, 2008; Lyon, 2011) in statistical mechanics, since objective probabilities seem to entail indeterminism necessarily.

The most well-known and successful reconciliation of chance and determinism in the context of statistical mechanics is defended by Loewer (2001). It is seldom recognized by interpreters, however, that there is no reconciliation in the sense of simultaneous compatibility between chance and determinism. The world cannot both be chancy and deterministic as a matter of metaphysical fact. As Lewis writes, "to the question of how chance can be reconciled with determinism, or to the question of how disparate chances can be reconciled with one another, my answer is: *it can't be done* (Lewis, 1986, 118). This is because chance entails indeterminism, the contrary of determinism. Thus, insofar as the probabilities of statistical mechanics and quantum mechanics are objective, these theories are indeterministic theories. Loewer's account actually shows us how deterministic laws can co-exist with indeterministic laws within a theory. The source of all probabilities in statistical mechanics, according to Loewer, is in an initial chance distribution over microscopic states of affairs. After the initial time these states of affairs evolve deterministically. Note that although for almost all times evolution is deterministic, it is not so at all times. There is an initial chance event, which is where the indeterminism of the theory appears. A deterministic theory is, recall, a theory whose laws are deterministic, not a theory whose laws are mostly deterministic or operate deterministically for almost all times.

Loewer's account is also presented in terms of Humean chances, so he does not believe

these chances and laws actually exist. According to the modern Humean, they merely are the result of the best systematizations of the occurrent facts, in keeping with Lewis's "best systems account" of laws and chances. Demarest, however, offers a small tweak to Loewer's Humean account by invoking a "robustly metaphysical account of chance" (Demarest, 2016, 256). She claims that such chances are compatible with determinism, and indeed they are when, as said, compatibility is understood to pertain to the co-existence of indeterministic and deterministic laws in a single theory—which, however, do not operate at the same time.[2]

Demarest's central claims are that this initial chance event exists and that we have good evidence for it. I dispute these claims in the remainder of the paper.

To begin, it is not so clear what exactly Demarest takes the evidence for the initial chance event to be. She does contrast the evidential position of her view with the Humean view of Loewer, claiming that, "for the Humean, the statistical patterns in the world are not evidence of an initial chance event" (Demarest, 2016, 261)—presumably this is so because Humeans reject the metaphysics of chance for the usual Humean reasons. One might suppose, then, that she believes that statistical patterns in the world are evidence of an initial chance event for all those who do not share the Humeans ontological worries. Let us accept, for the moment then, that statistical patterns may be *some* evidence for the existence of chances, for it is difficult to see what other evidence there might be for an initial chance event. In that case, on what grounds might we say that statistical patterns are good evidence for initial chances? I consider a series of three salient contrast classes.

First, do statistical patterns in data provide good evidence for indeterministic (i.e. chancy) theories *rather than deterministic theories*? It would seem that the answer is: not necessarily. (Werndl, 2009), for example, argues for the observational equivalence of indeterministic theories and deterministic theories. If one could contrive a fully deterministic theory that reproduces the same statistical patterns of the relevant phenomena observed in nature, then it would seem that such patterns provide no better evidence for the indeterministic theory than the deterministic one. However, since the theories under discussion, statistical mechanics and quantum mechanics, are generally characterized as indeterministic, let us flag but set aside the possibility of fully deterministic alternatives to them.

So, second, do statistical patterns provide good evidence for initial chances *rather than non-initial chances*? It would seem that the answer is firmly: no. There is a variety of ways one could implement chances into a probabilistic theory like statistical mechanics. All one must do, as Loewer shows us by example, is neatly separate when the indeterministic laws are operative and when the deterministic laws are operative. Loewer chooses to locate all the indeterminism in one place—the initial time—but one could equally locate it at another time, at many times, or even all times. Statistical mechanics does not wear its interpretation on its sleeve, just as quantum mechanics does not decide between solutions of the measurement problem, whether initial chances as in Bohmian mechanics or collapse

[2]Still, it is worth emphasizing that her claim that her account applies to deterministic worlds is false, for chancy worlds are not deterministic.

dynamics as in GRW (discrete time collapses) or CSL (continuous collapses). Unless there are evidential reasons to favor one implementation of indeteterministic probabilities over the others, there is not good evidence for an initial chance event. Certainly statistical patterns in nature will not do so.

Third, do statistical patterns provide good evidence for "robustly metaphysics" chances *rather than Humean chances*? It seems as if this might Demarest's intended contrast class, since much of the discussion in the paper concerns the Humean account. I will have something to say about the relative merits of Demarest's non-Humean account and Loewer's Humean account at the end of the next section. In any case though, it does not seem as if statistical patterns decide the matter in Demarest's mind, for she repeatedly demurs in the face of Humean responses to the considerations she raises, claiming only to offer an alternative "for philosophers who are antecedently sympathetic to governing laws of nature or powerful properties" (Demarest, 2016, 261-2). She finds it "plausible to think of the universe as having an initial state and as producing subsequent states in accordance with the laws of nature (some of which may be chancy)" (Demarest, 2016, 261). Such metaphysical intuitions are not grounded on observations of statistical patterns. Statistical patterns do not have any evidential bearing on the metaphysical dispute between the Humean and non-Humean.

Therefore, based on my canvassing of relevant alternatives, I conclude that we in fact do not have good evidence for an initial chance event, where evidence is interpreted in terms of statistical patterns (or in any usual sense of the term "evidence"). At best we have a motivation to attend to indeterministic theories when our evidence displays statistical patterns. It is another matter entirely to decide how to implement probabilities in that theory.

That said, Demarest's reasoning could be interpreted at points as invoking explanatory considerations as justification for the initial chance interpretation. Insofar as one considers "what justifies" as constituting evidence, perhaps these explanatory considerations should be counted as evidence.[3] Nevertheless, it does not look, on the face of it, like we have good evidence for an initial chance event still. Repeating the three cases considered before: deterministic and chancy theories can both serviceably explain statistical evidence; alternative implementations of chance in interpretations of indeterministic theories explain statistical evidence equally well; Humean and non-Humean metaphysics each render a story for how statistical patterns come about (merely subjective intuitions notwithstanding). Without explicit explanatory reasons to prefer one of these alternatives to the other, reasons lacking in Demarest's argument, good evidence (in this wider sense) for an initial chance event remains elusive.

[3]There are obvious dangers with going to far in this direction. Suppose that the Supreme Being explains all. Then it would appear that we have very good evidence of Its existence, which is obviously absurd.

3 Chance and Determinism in Systems of the World

In the previous section I gave reasons to doubt Demarest's claims about an initial chance event and our evidence for it. I disputed especially that we have evidence for it and did so by comparing it to alternatives of three different kinds. In the first case I characterized the issue (in part) as a matter of theory choice, namely of choosing between an indeterministic and deterministic theory. In the second case I characterized the issue as a matter of theory interpretation, namely of interpreting between different ways of implementing probability in a theory that does not decide one way or another on how this must be done. In the third case I characterized the issue as a matter of metaphysics, namely of deciding between the ontological status of chances.

In this section I consider more broadly whether there are any reasons to favor Demarest's interpretation, in particular in the sense of the just given second characterization of the issue. The question is whether the world should be thought to have an initial chance event, when one might consider that it is chancy in various other ways, e.g. its laws of evolution themselves are always probabilistically indeterministic.

First of all, it is worth mentioning that from the point of view given by the contemporary standard model of cosmology this question is moot. The so-called ΛCDM model, a development of the older standard big bang model, is a model of the general theory of relativity, a theory which makes use of no probabilities at all in its basic description of gravitating systems (including the universe). In this different sense it is also true that the universe never had a chance.

Demarest is not particularly interested in cosmology or the universes of general relativity however. She is concerned with probabilistic theories like classical statistical mechanics and quantum mechanics as applied to the world at large. We should, that is, imagine a statistical mechanical universe or a quantum mechanical universe (never minding that no concrete such model exists in physics that describes our universe) as a conceptual possibility when asking metaphysical questions about the world. Given the different ways of implementing probabilities in such a universe, we should ask whether one way is preferable to the others.

I should point out that this is not Demarest's question, for she explicitly restricts attention to "deterministically evolving worlds". Of course these worlds are not actually deterministic so long as the probabilities involved are chances. Nevertheless, unaffected by that fact is one of her central points: "that positing just one initial chance event can justify the usefulness and explain the ubiquity of nontrivial probabilities to epistemic agents like us, even if there are no longer any chance events in our world"(Demarest, 2016, 249). I say: so can a lot of other ways of conceiving chance in these theories. It is therefore necessary to compare them if we are to take Demarest's (and Loewer's) account seriously.

For present purposes, I am happy to agree with Demarest that the initial chance account can indeed justify and explain nontrivial probabilities used to describe subsystems of the universe.[4] But is it a good explanation? Is it worth believing?

[4]Notwithstanding pressure to move in this "global" direction in statistical mechanics (Callender, 2011)

The initial chance account invites the oft-invoked (in cosmology) picture of the (blind and unskilled) Creator throwing a dart (Wald, 2006, 396) or pointing a pin (Penrose, 1989, 442) at the set of possible universes, thereby picking out the initial conditions of the universe. That such pictures are intended as pejorative jabs at dubious metaphysics is plain. A mere picture is hardly an objection, of course, so what is it that seems problematic about initial chances for the universe? Could it not be the best cosmogonical story of our universe, that is, that a matter of chance determined its actualization out of a vast range of possibilities that could have been actualized had only their sisal been struck?

Intuition suggests that this just is not a serious, satisfying story for how the world could be. The probabilities of events in the actual world would derive ultimately from the probabilities for the actualization of our world. But why should we not just assume that the world started in the state that it did, with probability one or with certainty? Presumably the response of the initial chance advocate is that in that case we would lose the justification and explanation of subsystem probabilities. Yet is there anything to lose, if this metaphysical explanation is epistemically untrustworthy? How can we come to know these ultimate probabilities of other worlds? Is the metaphysical story sufficiently complete even? How could the probabilities of other worlds matter for what happens in *our* world?

I am willing to grant that these questions do have some answer, for what strikes me as a more serious difficulty is the following. Insofar as they are objective and justified, the probabilities agents like us use for specific events in subsystems of the world must be epistemic probabilities. On Demarest's (and Loewer's) account all such epistemic probabilities derive from initial epistemic probabilities for different initial conditions of the world. How is it that these probabilities obtain their needed objectivity and justification, and hence explanatory power? According to Demarest it is because they accord with the actual chances. However, what has one achieved by invoking "actual chances" at this stage? Although these chances do not merely have a *virtus dormitiva* per se, "just so" stories like this surely make the explanatory credentials of chances suspect. Does one dare invoke a transcendental argument or thump the realist table to defend their objectivity?

If we were somehow forced to adopt the initial chance explanation of epistemic probabilities, then we might swallow whatever dubious metaphysics attendant to it. If there were reasonable alternatives, however, should we not prefer them? And indeed there are other interpretive options available. Locating the chances at another time (or even "outside the universe") constitutes one set of possibilities, but they obviously suffer from the same awkwardness as the initial chance account. Another is based on the idea that chancy behavior occurs at discrete time intervals. One finds this idea in the orthodox Copenhagen and other collapse interpretations of quantum mechanics for example. One might be uneasy with the invocation of chancy behavior at potentially ill-defined times in such interpretations, and even with their postulation of two dynamical laws of nature, a deterministic one and an indeterministic one (although it is a feature of the initial chance account as well). However one at least avoids a commitment to chance figuring into

(and quantum mechanics) in order to justify and explain probabilities in subsystems of the universe, serious reservations about whether doing so is itself justified are advanced by, inter alia, Earman (2006).

cosmogenesis and also the questionable leap to objectivity in agential probabilities, since chances in these interpretations are physical processes that happen within the universe, whether as part of the general evolution of the universe or tied to the evolution of individual systems.

Another possibility is suggested by continuing this line of thought, i.e. of spreading chanciness out further in time. Instead of chancy behavior at discrete intervals, why not suppose that it occurs continuously? In quantum mechanics this idea is implemented in some interpretations, such as continuous spontaneous localization, and in statistical mechanics there are various stochastic dynamics approaches. Advantages of this idea are that one has a single law of evolution, an indeterministic one, and, again, one does not make chanciness a matter of cosmogenesis. What disadvantage? To some that it makes the world rife with indeterminism. Yet who is afraid of indeterminism? It surely does not mean anything goes, nor does it threaten the possibility of knowledge of the world (although there are limits to what we can know). Besides, by accepting quantum mechanics (or even statistical mechanics) we have already let indeterminism in the door in physics.

When we look at the interpretations available for a world governed by probabilistic laws, in every case the alternatives to the initial chances view therefore appear preferable. Indeed, it would seem that only one who demands that the world be as deterministic as possible could favor the initial chances view, but it is hard to see what motivation there could be for that demand. I therefore conclude, in a final sense, that *the universe never had a chance*.

That said, I emphasize that this judgment applies only to the case where we treat the universe as a statistical mechanical system or quantum mechanical system. In other words, the world is the universe, our world-metaphysics is our universe-metaphysics. The considerations leading to this conclusion change shape somewhat when we confine the application of our theories to systems describable by those theories. The initial chance account is far less dubious when attached to individual statistical mechanical systems and not automatically to the universe at large. Indeed, it could well be that the initial conditions of similar systems are best treated as randomly distributed, for here we do have empirical evidence that this interpretation can be used to explain—unlike with the universe, where we have but one system.

There is, as noted, sometimes pressure to globalize our theories, especially in the case of statistical mechanics. If we ask what accounts for the randomness in initial conditions of a particular class of systems, it is natural to look at larger systems that contain them. If we find that these systems have random initial conditions, then we continue to expand our scope, ultimately reaching the "maximally inclusive entity whose parts are all the things that exist." This globalization of statistical mechanics is the kernel of the so-called imperialism of (Albert, 2000) and Loewer. If we are right to feel this pressure to interpret the world at large in the same terms as individual physical systems, then there is concomitant pressure to hold the same interpretive of chance in both cases. I have argued, however, that the intuitive considerations vary somewhat, at least with respect to the initial chance account. Is this reason to disfavor it in the case of individual systems? Or is our confidence in its applicability for individual systems sufficient to overcome any hesitation at

accepting it for the universe? My inclination is to answer "yes" and "no", but I offer no grounds for the preference here. I do believe that metaphysicians of science should care about considerations like this, however, having to do with the relation of subsystem and universe, for often enough what seems right in one context is questionable in the other.

I close this section with a brief comment on the relation of Loewer's and Demarest's accounts. As I argued above, empirical evidence and explanatory considerations do not favor one over the other, since they account for empirical evidence in essentially the same way. The central difference is whether chances are understood as reducible to other facts, hence not part of the fundamental ontology of the world, or as "robustly metaphysical", in which case they are. The problems Demarest mentions for the Humean view—past events may have nontrivial chances, the chance of an event depends on what one knows, worlds with identical frequencies cannot have different chances, etc.—are surely not problems when viewed properly through the Humean lens. However, whereas the problem I raise for the initial chance view, concerning the explanatory credentials and justification for the posit of initial chances, threatens Demarest's account, it will not worry the Humean of Loewer's stripe, for these initial chances do not exist for the Humean. Humean chances do not produce or generate any actual states of affairs. Of course one may raise the usual complaint against the Humean, that there is a circularity in the Humean account involving descriptions explaining themselves, and others besides. I do not care to enter into this debate here of course. I only wish to point out that my argument about how chance can fit into a cosmogonical worldview appears to give some reason to favor the Humean account in this particular context.

4 Conclusion

In this paper I considered whether we should think that the world had one chance, as claimed by Demarest. First I considered her claim that we have good evidence that an initial chance event occurred by contrasting it with relevant classes of alternatives. I argued that evidence neither favors a chancy theory over a chanceless theory, nor initial chances over other implementations of chances, nor metaphysically robust chances over Humean chances. I concluded, therefore, that we do not have good evidence to adopt the initial chance account.

I then considered whether there were other reasons to favor or disfavor the initial chance account. I argued that the dubious nature of worldly chances provides a strong impulse to look for other accounts that do not make chance a matter of cosmogenesis. The other implementations did not suffer from this defect, so I suggested that from a cosmogonical perspective they should be preferred. But the relation of the universe and its subsystems makes a demand to have a consistent interpretation. As the initial chance account looks favorable on the subsystem level (to many) and not on the universe's level (as I argued), there remains a significant metaphysical tension to be resolved.

References

Albert, D. (2000). *Time and Chance*. Cambridge, MA: Cambridge, MA: Harvard University Press.

Callender, C. (2011). The past histories of molecules. In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 83–113. Oxford: Oxford University Press.

Demarest, H. (2016). The universe had one chance. *Philosophy of Science 83*(2), 248–264.

Earman, J. (1986). *A Primer on Determinism*. Dordrecht: D. Reidel Publishing Company.

Earman, J. (1995). *Bangs, Crunches, Whimpers, and Shrieks*. Oxford: Oxford University Press.

Earman, J. (2006). The "past hypothesis": Not even false. *Studies in History and Philosophy of Modern Physics 37*, 399–430.

Earman, J. and J. Norton (1987). What price spacetime substantivalism? the hole story. *British Journal for the Philosophy of Science 38*, 515–525.

Harrison, E. (2000). *Cosmology: the science of the universe* (2nd ed.). Cambridge: Cambridge University Press.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Lewis, D. (1986). *Philosophical Papers*, Volume 2. Oxford: Oxford University Press.

Loewer, B. (2001). Determinism and chance. *Studies in History and Philosophy of Modern Physics 32*, 609–620.

Lyon, A. (2011). Deterministic probability: neither chance nor credence. *Synthese 182*, 413–432.

Norton, J. (2008). The dome: An unexpectedly simple failure of determinism. *Philosophy of Science 75*, 786–798.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

Schaffer, J. (2010). Monism: The priority of the whole. *The Philosophical Review 119*, 31–76.

Wald, R. (2006). The arrow of time and the initial conditions of the universe. *Studies in History and Philosophy of Modern Physics 37*, 394–398.

Werndl, C. (2009). Are deterministic descriptions and indeterministic descriptions observationally equivalent? *Studies in History and Philosophy of Modern Physics 40*, 232–242.

Winsberg, E. (2008). Laws and chances in statistical mechanics. *Studies in History and Philosophy of Modern Physics 39*, 872–888.

Wüthrich, C. (2011). Can the world be shown to be indeterministic after all? In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 365–389. Oxford: Oxford University Press.

Draft paper for the symposium *Mechanism Meets Big Data: Different Strategies for Machine Learning in Cancer Research* to be held at *PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association* (Seattle, WA; 1-4 Nov 2018).

MECHANISTIC MODELS AND THE EXPLANATORY LIMITS OF MACHINE LEARNING

Emanuele Ratti[1], University of Notre Dame

Ezequiel López-Rubio, Universidad Nacional de Educación a Distancia, University of Málaga

Abstract

We argue that mechanistic models elaborated by machine learning cannot be explanatory by discussing the relation between mechanistic models, explanation and the notion of intelligibility of models. We show that the ability of biologists to understand the model that they work with (i.e. intelligibility) severely constrains their capacity of turning the model into an explanatory model. The more a mechanistic model is complex (i.e. it includes an increasing number of components), the less explanatory it will be. Since machine learning increases its performances when more components are added, then it generates models which are not intelligible, and hence not explanatory.

1. INTRODUCTION

Due to its data-intensive turn, molecular biology is increasingly making use of machine learning (ML) methodologies. ML is the study of generalizable extraction of patterns from data sets starting from a problem. A problem here is defined as a given set of input variables, a set of outputs which have to be calculated, and a sample (previously input-output pairs already observed). ML calculates a quantitative relation between inputs and outputs in terms of a predictive model by learning from an already structured set of input-output pairs. ML is expected to increase its performances when the complexity of data sets increase, where complexity refers to the number of input variables and the number of samples. Due to this capacity to handle complexity, practitioners think that ML is potentially able to deal with biological systems at the macromolecular level, which are notoriously complex. The development of ML has been proven useful not just for the

[1] mnl.ratti@gmail.com

complexity of biological systems *per se*, but also because biologists now are able to generate an astonishingly amount of data. However, we claim that the ability of ML to deal with complex systems and big data comes at a price; *the more ML can model complex data sets, the less biologists will be able to explain phenomena in a mechanistic sense*.

The structure of the paper is as follows. In Section 2, we discuss mechanistic models in biology, and we emphasize a surprising connection between explanation and model complexity. By adapting de Regt's notion of pragmatic understanding (2017) in the present context, we claim that if a how-possibly mechanistic model can become explanatory, then it must be intelligible to the modeler (Section 2.2, 2.3 and 2.4). Intelligibility is the ability to perform precise and successful material manipulations on the basis of the information provided by the model about its components. The results of these manipulations are fundamental to recompose the causal structure of a mechanism out of a list of causally relevant entities. Like a recipe, the model must provide instructions to 'build' the phenomenon, and causal organization is fundamental in this respect. If a model is opaque to these organizational aspects, then no mechanistic explanations can be elaborated. By drawing on studies in cognitive psychology, we show that the more the number of components in a model increases (the more the model is complex), the less the model is intelligible, and hence the less an explanation can be elaborated.

Next, we briefly introduce ML (Section 3). As an example of ML application to biology, we analyze an algorithm called PARADIGM (Vaske et al 2010), which is used in biomedicine to predict clinical outcomes from molecular data (Section 3.1). This algorithm predicts the activities of genetic pathways from multiple genome-scale measurements on a single patient by integrating information on pathways from different databases. By discussing the technical aspects of this algorithm, we will show how the algorithm generates models which are more accurate as the number of variables included in the model increases. By variables, here we mean biological entities included in the model and the interactions between them, since those entities are modeled by variables in PARADIGM.

In Section 4 we will put together the results of Section 2 and 3. While performing complex localizations more accurately, we argue that an algorithm like PARADIGM makes mechanistic models so complex (in terms of the number of model components) that no explanation can be constructed. In other words, ML applied to molecular biology undermines biologists' explanatory abilities.

2. COMPLEXITY AND EXPLANATIONS IN BIOLOGY

The use of machine learning has important consequences for the explanatory dimension of molecular biology. Algorithms like PARADIGM, while providing increasingly accurate localizations, challenge the explanatory abilities of molecular biologists, especially if we assume the account of explanation of the so-called mechanistic philosophy (Craver and Darden 2013; Craver 2007; Glennan 2017). In order to see how, we need to introduce the notion of mechanistic explanation, and its connection with the notion of intelligibility (de Regt 2017).

2.1 Mechanistic explanations

Molecular biology's aim is to explain how phenomena are produced and/or maintained by the organization instantiated by macromolecules. Such explanations take the form of mechanistic descriptions of these dynamics. As Glennan (2017) succinctly emphasizes, mechanistic models (often in the form of diagrams complemented by linguistic descriptions) are vehicles for mechanistic explanations. Such explanations show how a phenomenon is produced/maintained and constituted by a mechanism – mechanistic models explain by explaining *how*. As Glennan and others have noticed, a mechanistic description of a phenomenon looks like what in historical narrative is called *causal narrative*, in the sense that it "describes sequences of events (which will typically be entities acting and interacting), and shows how their arrangement in space and time brought about some outcome" (Glennan 2017, p 83). The main idea is that we take a set of entities and activities to be causally relevant to a phenomenon, and we explain the phenomenon by showing how a sequence of events involving the interactions of the selected entities produces and/or maintains the explanandum. In epistemic terms, it is a

matter of showing a chain of inferences that holds between the components of a model (e.g. biological entities). Consider for instance the phenomenon of restriction in certain bacteria and archaea (Figure 1). This phenomenon has been explained in terms of certain entities (e.g. restriction and modification enzymes) and activities (e.g. methylation). Anytime a bacteriophage invades one of these bacteria or archaea (from now on *host cells*), host cells stimulate the production of two types of enzymes, i.e. a restriction enzyme and a modification enzyme. The restriction enzyme is designed to recognize and cut specific DNA sequences. Such sequences, for reasons we will not expose here[2], are to be found in the invading phages and/or viruses. Hence, the restriction enzyme destroys the invading entities by cutting their DNA. However, the restriction enzyme is not able to distinguish between the invading DNA and the DNA of the host cell. Here the modification enzyme helps, by methylating the DNA of the host cell at specific sequences (the same that the restriction enzyme cuts), thereby preventing the restriction enzyme to destroy the DNA of the host cell. The explanation of the phenomenon of restriction is in terms of a narrative explaining how certain entities and processes contribute to the production of the phenomenon under investigation. The inferences take place by thinking about the characteristics of the entities involved, and how the whole functioning of the system can be recomposed from entities themselves.

[2] See for instance (Ratti 2018)

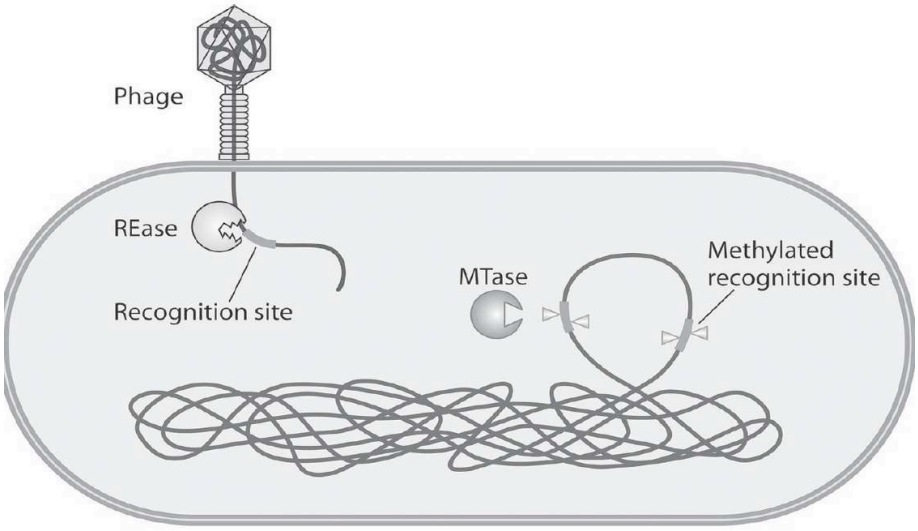


Figure 1. Mechanistic model of restriction. A phage enters a bacterium cell and sequences of its DNA are cleaved by a restriction enzyme (REase). Simultaneously, a modification enzyme (MTase) methylates a specific sequence in the DNA of host so that the restriction enzyme does not cleave the genome of the host too. Original figure taken from (Vasu and Nagaraja 2013).

2.2. Complexity of mechanistic models

Despite the voluminous literature on mechanistic explanation, there is a connection between models, *in fieri* explanations and the modeler that has not been properly characterized. In particular, mechanistic models should be intelligible to modelers in order to be turned into complete explanations. Craver noticed something like that when he states that his ideal of completeness of a mechanistic description (in terms of molecular details) should not be taken literary, but completeness always refer to the particular explanatory context one is considering. The reason why literary completeness is unattainable is because complete models will be of *no use* and completely *obscure* to modelers; "such descriptions would include so many potential factors that they would be *unwieldy for the purpose of prediction and control and utterly unilluminating to human beings*" (2006, p 360, emphasis added).

We rephrase Craver's intuitions by saying that *how-possibly models cannot be turned into adequate explanations if they are too complex*. We define complexity as *a function of the number of entities and activities (i.e. components of the model) that have*

to be coordinated in an organizational structure in the sense specified by mechanistic philosophers. This means that no agent can organize the entities and/or activities localized by highly complex models in a narration that rightly depicts the organizational structure of the *explanandum*. Therefore, very complex models which are very good in localization cannot be easily turned into explanations. Let us show why complex models cannot be turned into explanatory models in the mechanistic context.

2.3 Intelligibility of mechanistic models

The idea that agents cannot turn highly complex mechanistic models into explanations can be made more precise by appealing to the notion of *intelligibility* (de Regt 2017).

By following the framework of models as mediators (Morgan and Morrison 1999), de Regt argues that models are the way theories are applied to reality. Similar to Giere (2010), de Regt thinks that theories provide principles which are then articulated in the form of models to explain phenomena; "[t]he function of a model is to represent the target system in such a way that the theory can be applied to it" (2017, p 34). He assumes a broad meaning of explanation, in the sense that explanations are arguments, namely attempts to "answer the question of why a particular phenomenon occurs or a situation obtains (…) by presenting a systematic line of reasoning that connects it with other accepted items of knowledge" (2017, p 25). *Ça va sans dire*, arguments of the sort are not limited to linguistic items[3]. On this basis, de Regt's main thesis is that a *condition sine qua non* to elaborate an explanation is that the theory from which it is derived must be intelligible.

In de Regt's view, the intelligibility of a theory (*for scientists*) is "[t]he value that scientists attribute to the cluster of virtues (…) that facilitate the use of the theory for the construction of models" (p 593). This is because an important aspect of obtaining explanations is to derive models from theories, and to do that a scientist must use the theories. Therefore, if a theory possesses certain characteristics that make it easier to be used by a scientist, then the same scientist will be in principle more successful in deriving explanatory models. In (2015) de Regt extends this idea also to models in the sense that "understanding consists in being able to use and manipulate the model in order to make

[3] Mechanistic explanations are arguments, though not of a logical type

inferences about the system, to predict and control its behavior" (2015, p 3791). If for some reasons models and theories are not intelligible (to us), then we will not be able to develop an explanation, because we would not know how to use models or theories to elaborate one.

This idea of intelligibility of models and its tight connection with scientific explanation, can be straightforwardly extended to mechanistic models. Intelligibility of mechanistic models is defined by the way we *successfully* use them to explain phenomena. But how do we use models (mechanistic models in particular), and for what? Please keep in mind that whatever we do with mechanistic models, it is with explanatory aims in mind. Anything from predicting, manipulating, abstracting, etc is because we want an explanation. This is a view shared both by mechanistic philosophers but by de Regt as well, whose analysis of intelligibility is in explanatory terms.

First, highly abstract models can be used to build more specific models, as in the case of schema (Machamer et al 2000; Levy 2014). A schema is "a truncated abstract description of a mechanism that can be filled with descriptions of known component parts and activities" (Machamer et al 2000, p 16). For instance, consider the model of transcription. This model can be highly abstract where 'gene' stands for any gene, and 'transcription factor' stands for any transcription factor. However, we can instantiate such a schema in a particular experimental context by specifying which gene and which transcription factors are involved. The idea is that biologists, depending on the specific context they are operating, can instantiate experiments to find out which particular gene or transcription factor is involved in producing a phenomenon at a given time.

Next, mechanistic models can be used in the context of the *build-it test* (Craver and Darden 2013) with confirmatory goals in mind. Since mechanistic explanations may be understood as recipes for construction, and since recipes provide instructions to use a set of ingredients and instruments to produce something (e.g. a cake), then mechanistic models provide instructions to build a phenomenon or instructions to modify it in controlled ways because, after all, they tell us about the internal division of labor between entities causally relevant to producing or maintaining phenomena. This is in essence the build-it test as a confirmation tool; by modifying an experimental system on the basis of the 'instructions' provided by the model that allegedly explains such a phenomenon, we

get hints as to how the model is explanatory. If the hypothesized modifications produce in the 'real-world' the consequences we have predicted on the basis of the model, then the explanatory adequacy of the model is corroborated. The more the modifications suggested are precise, the more explanatory the model will be[4]. A first lesson we can draw is that *if a mechanistic model is explanatory, then it is also intelligible*, because it is included in the features of being explanatory mechanistically the fact that we can use the model to perform a build-it test.

The build-it test is also useful as a *tool to develop* explanations. Consider again the case of restriction in bacteria and a how-possibly model of this phenomenon based on a few observations. Let's say that we have noticed that when phages or viruses are unable to grow in specific bacteria, such bacteria also produce two types of enzymes. We know that the enzymes, the invading phages/viruses and restriction are correlated. The basic model will be as follows; anytime a phage or a virus invade a bacterium, these enzymes are produced, and hence the immune system of the bacterium must be related to these enzymes. We start then to instantiate experiments on the basis of this simple model. Such a model suggests that these enzymes must do something to the invading entities, but that somehow modify the host cell as well. Therefore, the build-it test would consist in a set of experiments to stimulate and/or inhibit these entities to develop our ideas about the nature of their causal relevance and their internal division of labor. *In fieri* mechanistic models suggest a range of instructions to 'build' or 'maintain' phenomena. These instructions are used to instantiate experiments to refine the model and make it explanatory. This is an example of what Bechtel and Richardson would call *complex localization* (2010, Chapter 6), and it is complex because the strategy used to explain the behavior of a system (immune system of host cells) is heavily constrained by empirical results of lower-levels. The how-possibly model affords a series of actions leading to a case of complex localization, when "constraints are imposed, whether empirical or theoretical, they can serve simultaneously to vindicate the initial localization and to develop it into a full-blooded mechanistic explanation" (Bechtel and Richardson 2010, p 125). Therefore, *if a how-possibly model can be turned into an explanatory model, then it*

[4] Please note that such a test, when involving adequate mechanistic explanations, is also the preferred way to teach students in text books, or also a way to provide instructions to reproduce the results of a peer-reviewed article

is intelligible, because the way we turn it into an explanatory model is by instantiating build-it tests.

A mechanistic model is therefore intelligible either when (a) it is a schema and we can instantiate such a model in specific contexts, or (b) when it affords a series of built-it test which are used either to corroborate its explanatory adequacy, or to make it explanatory. About (b), it should be noted that if we consider a mechanistic model as a narrative, then the model will be composed of a series of steps which influence each other in various ways. *Being able to use a model means being able to anticipate what would happen to other steps if I modify one step in particular*. This is not a yes/no thing. The model of restriction-modification systems is highly intelligible, because I know that if I prevent the production of modification enzymes I simultaneously realize that the restriction enzyme will destroy the DNA of the host cell. However, more detailed models will be less intelligible, because it would be difficult to simultaneously anticipate what would happen at each step by modifying a step in particular.

2.4 Recomposing mechanisms and intelligibility

In the mechanistic literature, the process of developing an explanatory model out of a catalogue of entities that are likely to be causally relevant to a phenomenon is called *recomposition of a mechanism* and it usually happens after a series of localization steps.

To recompose a mechanism, a modeler must be able to identify causally relevant entities and their internal division of labor. The idea is not just to 'divide up' a given phenomenon in tasks, but also a given task in subtasks interacting in the overall phenomenon, as it happens in complex localization (Bechtel and Richardson 2010). In the simplest case, researchers assume linear interactions between tasks, but there may be also non-linear or more complex type of interactions.

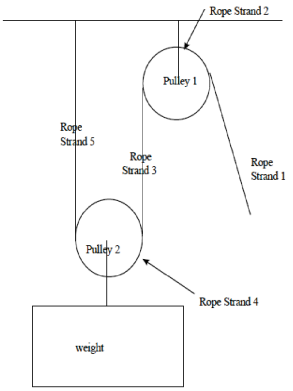
These reasoning strategies are usually implemented by thinking about these dynamics with the aid of *diagrams*. Diagrammatic representations usually involve boxes standing for entities (such as genes, proteins, etc) and arrows standing for processes of various sorts (phosphorylation, methylation, binding, releasing, etc). Therefore, biologists recompose mechanisms as mechanistic explanations by thinking about these diagrams,

and they instantiate experiments (i.e. built-it test) exactly on the basis of such diagrammatic reasoning.

Cognitive psychology and studies of scientific cognition have extensively investigated the processes of diagrammatic reasoning (Hegarty 2000; 2004; Nersessian 2008). Moreover, empirical studies have emphasized the role of diagrams in learning and reasoning in molecular biology (Kindfield 1998; Trujillo 2015). In these studies, diagrammatic reasoning is understood as a "task that involves inferring the behavior of a mechanical system from a visual-spatial representation" (Hegarty 2000, p 194). Hegarty refers to this process as *mental animation*, while Nersessian (2008) thinks about this as an instantiation of *mental modelling*. This is analogous to thinking about mechanistic models as narratives, namely being able to infer how a course of events, decomposed into steps, may change if we change one step in particular. Mental animation is a process of complex visual-spatial inference. Limits and capabilities of humans in such tasks depend on the cognitive architecture of human mind[5]. What Hegarty has found is that mental animation is *piecemeal*, in the sense that human mind does not animate the components of a diagram in parallel, but rather infer the motion of components *one by one*. This strategy has a straightforward consequence; in order to proceed with animating components, we should store intermediate results of inferences drawn on previous components. Due to the limitations of working memory (WM), people usually store such information on external displays. Hegarty has provided evidence that diagrammatic reasoning is bounded to WM abilities. The more we proceed in inferring animation on later components, the more the inferences on earlier components degrade (see for instance Figure 2); "as more components of the system are 'read into' spatial working memory, the activation of all items is degraded, so that when later components are in, there is not enough activation of the later components to infer their motion" (Hegarty 2000, p 201).

[5] On this, I rely on the framework assumed by the cognitive-load theory (Paas et al 2010)

a



b

Cycle	Action Taken
1	Rope Strand 1 is inspected (simulation is given knowledge that this rope strand is being pulled)
2	Rope Strand 1 is animated (i.e. its direction of motion is inferred)
3	Rope Strand 2 is inspected
4	Rope Strand 2 is animated
5	Rope Strand 3 is inspected Pulley 1 is inspecte
6	Rope Strand 3 is animated Pulley 1 is animated
7	Rope Strand 4 is inspected
8	Rope Strand 4 is animated
9	Rope Strand 5 is inspected Pulley 2 is inspected
10	Rotation of Pulley 2 is inferred Translation of Pulley 2 is inferred
11	Weight is inspected
12	Weight is animated

Figure 2. (a) Example of diagram of a simple pulley system that can be mentally animated (b) Description of typical actions that can be one by one to animate the pulley system. Both figures taken from Hergaty (2000)

The actual limit of our cognitive architecture on this respect may be debated, and it is an empirical issue. The important point is that *no matter our external displays*, for very large systems (such as Figure 3) it is very unlikely that human cognition will be able to process all information about elements interactivity. This is because by animating components one-by-one, even if we use sophisticated instruments such computer simulations, still inferences on earlier components will degrade. This means that build-it tests will be very ineffective, if not impossible. In terms of narratives, recipes and mechanistic models, this means that for large mechanistic diagrams with many model components, no human would be able to anticipate the consequences of modifying a step in the model for all the other steps of the model, even if a computer simulation shows that the phenomenon can be possibly produced by the complex model. The computer simulation may highlight certain aspects (as Bechtel in 2016 notes), but the model is not intelligible in the sense required by mechanistic philosophy. *If the model is not intelligible in this way, then it cannot be possibly turned into an explanation*.

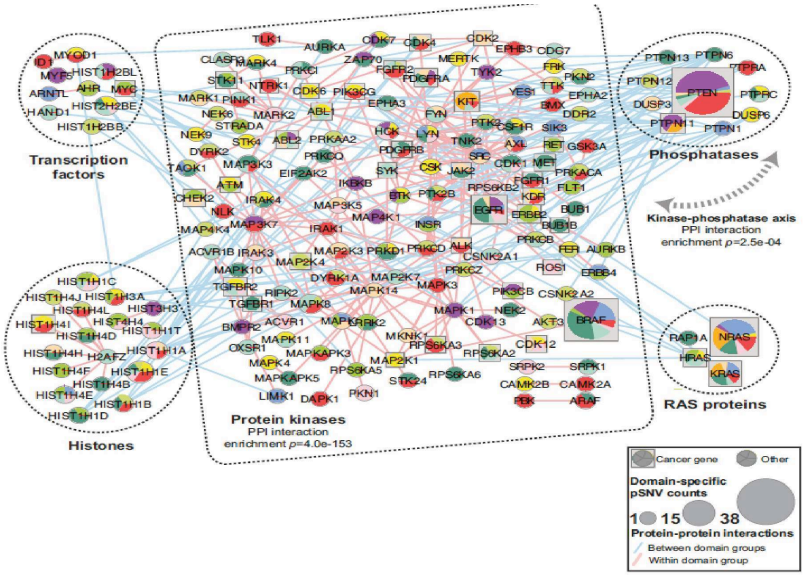


Figure 3. Network of interactions of proteins with significant enrichment of phosphorylation-related single nucleotide variations. Phosphorylation is a central post-translational modification in cancer biology. Authors are not trying to recompose the mechanism that from phosphorylated proteins (nodes) lead to a tumor phenotype, but rather to identify the magnitude of the impact of this process on cancer genes. Figure taken from (Reimand et al 2013)

The results of Hegarty's research suggest that when mechanistic models are concerned, strategies of localization are effective (in terms of explanatory potential) only when a limited number of model components are actually identified. The number may increase if we use computer simulations. However, for very large amounts of model components (such as Figure 3) recomposition is just impossible for humans, because inferences on the role of components in the causal division of labor of a phenomenon will degrade to make place for inferences about other components. This of course holds only if we have explanatory aims in mind.

To summarize, in section 2 we have made three claims:

1. If a how-possibly model can be turned into an explanation, then it is intelligible
2. If a model is not intelligible, then it cannot become explanatory
3. Complex models are a class of non-intelligible models

3. MACHINE LEARNING AND LOCALIZATIONS

Machine learning (ML) is a subfield of computer science which studies the design of computing machinery that improves its performance as it learns from its environment. A ML algorithm extracts knowledge from the input data, so that it can give better solutions to the problem that it is meant to solve. This learning process usually involves the automatic construction and refinement of a model of the incoming data. In ML terminology, a model is an information structure which is stored in the computer memory and manipulated by the algorithm.

As mentioned before, the concept of 'problem' in ML has a specific meaning which is different from other fields of science. A ML problem is defined by a set of input variables, a set of output variables, and a collection of samples which are input-output pairs. Solving a problem here means finding a quantitative relation between inputs and outputs in the form of a predictive model, in the sense that the algorithm will be used to produce a certain output given the presence of a specific input.

3.1 The PARADIGM algorithm

ML has been applied in the molecular sciences in many ways (Libbrecht and Noble 2015). Especially in cancer research[6], computer scientists have created and trained a great deal of algorithms in order to identify entities that are likely to be involved in the development of tumors, how they interact, to predict phenotypes, to recognize crucial sequences, etc (see for instance Leung et al 2016).

As a topical example of ML applied to biology, we introduce an algorithm called PARADIGM (Vaske et al 2010). This algorithm is used to infer how genetic changes in a patient influence or disrupt important genetic pathways underlying cancer progression. This is important because there is empirical evidence that "when patients harbor genomic alterations or aberrant expression in different genes, these genes often participate in a common pathway" (Vaske et al 2010, p i237). Because pathways are so large and biologists cannot hold in their mind the entities participating in them, PARADIGM integrate several genomic datasets – including datasets about interactions between genes and phenotypic consequences – to infer molecular pathways altered in patients; it predicts

[6] See for instance The Cancer Genome Atlas at https://cancergenome.nih.gov

whether a patient will have specific pathways disrupted given his/her genetic mutations.

The algorithm is based on a simplified model of the cell. Each biological pathway is modeled by a graph. Each graph contains a set of nodes, such that each node represents a cell entity, like a mRNA, a gene or a complex. A node can be only in three states (i.e. activated, normal or deactivated). The connections among nodes are called factors, and they represent the influence of some entities on other entities. It must be noticed that the model does not represent why or how these influences are exerted. Only the sign of the influence, i.e. positive or negative, is specified.

The model specifies how the expected state of an entity must be estimated. The entities which are connected by positive or negative factors to the entity at hand cast votes which are computed by multiplying +1 or –1 by the states of those entities, respectively. In addition to this, there are 'maximum' and 'minimum' connections to cast votes which are the maximum or the minimum of the states of the connected entities, respectively. Overall, the expected state of an entity is computed as the result of combining several votes obtained from the entities which are connected to it. Such a voting procedure can be associated to localizations (i.e. whether a node is activated or not), but hardly to biological explanations.

The states of the entities can be hidden, i.e. they can not be directly measured on the patients, or observable. The states of the hidden variables must be estimated by a probabilistic inference algorithm, which takes into account the states of the observed variables and the factors to estimate the most likely values of the hidden variables. Here it must be pointed out that this algorithm does not yield any explanation about the computed estimation. Moreover, it could be the case that the estimated values are not the most likely ones, since the algorithm does not guarantee that it finds the globally optimum solution.

The size of the model is determined by the number of entities and factors that the scientist wishes to insert. A larger model provides a perspective of the cell processes which contains more elements, and it might yield better predictions. This means that the more components the model has, the better the algorithm will perform. In biological terms, the larger the model, the more precise *complex localizations* the algorithm will identify, in particular by pointing more precisely towards pathways that are likely to be

disrupted in the patient with more information about the state of gene activities, complexes and cellular processes. Importantly, PARADIGM does not infer new genetic interactions, but it just helps identifying those known interaction in a new data set. It is completely supervised, in the sense that "[w]hile it infers hidden quantities (…), it makes no attempt to infer new interactions not already present in an NCI [National Cancer institute database] pathway" (Vaske et al 2010, p i244).

4 COMPLEX MODELS AND MECHANISTIC EXPLANATIONS

Before unwinding our conclusions, let me recall the results of Section 2 very briefly:

1. If a how-possibly model can be turned into an explanation, then it is intelligible[7].
2. If a model is not intelligible, then it cannot become explanatory
3. Complex model (in the sense explained in 2.2) are not intelligible

What does this have to do with PARADIGM? It is important to emphasize what we have pointed out in Section 3.1, namely that an algorithm like PARADIGM is more efficient when working with more components. If we think about models generated by algorithms such as PARADIGM in mechanistic terms, this means that the algorithm provides more precise complex localizations, because more entities that are likely to be causally relevant to a phenomenon are identified, and the information about the probability of a pathway being disrupted in a patient will be more precise. However, the models will be more complex, and they will be decreasingly intelligible. This is because the final model will count an elevated number of components, and recomposing these components into a full-fledged mechanistic explanation of how a tumor is behaving will be cognitively very difficult; the inferences about the behavior of components are not run in parallel, but one by one, and once we proceed in inferring the behavior of a component on the basis of the behavior of another component, other inferences will degrade, as Hegarty's studies have shown. In the ideal situation, PARADIGM will generate unintelligible models:

[7] Remember: A mechanistic model *x* is intelligible to a modeler *y* if *y* can use the information about the components of *x* to instantiate so-called 'build-it test'. Such tests are performed on how-possibly models to turn them into explanatory models by obtaining information on how to recompose a phenomenon (i.e. by showing how a list of biological entities are organized to produce a phenomenon).

4. Algorithms such as PARADIGM generate models which are not intelligible because such models are too complex
5. Because of 2, 3 and 4, complex models generated out of algorithms like PARADIGM cannot become explanations

This means that when we use algorithms such as PARADIGM to cope with the complexity of biological systems, we successfully handle big data sets, but such a mastery comes at a price. Using ML in molecular biology means providing more detailed localizations, but we also lose explanatory power, because no modeler will be able to recompose the mechanism out of a long list of entities.

This implies that, in the mechanistic epistemic horizon, the central role assigned to explanations should be reconsidered when contemporary molecular biosciences are concerned. As Bechtel has also emphasized in the context of computational models in mechanistic research (2016), such tools are useful to show whether some entities are likely to be involved in a particular phenomenon or suggest alternative hypotheses about the relation between certain entities. However, providing fully-fledged mechanistic explanations is another thing. It is the same with algorithms of ML; we identify more entities likely to be involved in a mechanism, we may even find out that entities involved in specific process may be connected with entities involved in other processes (via for instance Gene Ontology enrichments), but we cannot recompose a mechanism out of a list of hundreds of entities. In fact, we come to value different epistemic values, and *explanatory power is not one of them*. This somehow implies also a shift in the way scientific articles are organized; if in 'traditional' molecular biology evidence converges towards the characterization of a single mechanism, in data-intensive biology we make a list of entities that can be involved in a phenomenon, but we do not necessarily connect those entities mechanistically (Alberts 2012). Another strategy (Krogan et al 2015) – though motivated more by biologically rather than cognitive reasons – is to abstract from macromolecular entities and consider only aggregates of them in the form of networks; whether establishing network topology is providing a mechanistic explanation remains an open question.

REFERENCES

Alberts, B. (2012). The End of "Small Science"? *Science*, *337*(September), 1230529.

Bechtel, W. (2016). Using computational models to discover and understand mechanisms. *Studies in History and Philosophy of Science Part A*, *56*, 113–121.

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity - Decomposition and Localization as Strategies in Scientific Research*. Cambridge, Massachusetts, and London, England: The MIT Press.

Craver, C. (2007). *Explaining the Brain - Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. Chicago: The University of Chicago Press.

De Regt, H. (2017). *Understanding Scientific Understanding*. Oxford: Oxford University Press.

de Regt, H. W. (2015). Scientific understanding: truth or dare? *Synthese*, *192*(12), 3781–3797. http://doi.org/10.1007/s11229-014-0538-7

Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese*, *172*(2), 269–281.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford University Press.

Hegarty, M. (2000). Capacity Limits in Mechanical Reasoning. In M. Anderson, P. Cheng, & V. Haarslev (Eds.), *Diagrams 2000* (pp. 194–206). Springer-Verlag.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A., & Ideker, T. (2015). The Cancer Cell Map Initiative: Defining the Hallmark Networks of Cancer. *Molecular Cell*, *58*(4), 690–698.

Levy, A. (2014). What was Hodgkin and Huxley's achievement? *British Journal for the Philosophy of Science*, *65*(3), 469–492.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about Mechanisms. *Philosophy of Science*, (67), 1–25.

Morrison, M., & Morgan, M. (1999). Models as mediating instruments. In M. Morrison & M. Morgan (Eds.), *Models as Mediators*. Cambridge University Press.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: The MIT Press.

Ratti, E. (2018). "Models of" and "models for": On the relation between mechanistic models and experimental strategies in molecular biology. *British Journal for the Philosophy of Science*.

Reimand, J., Wagih, O., & Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Scientific Reports*, *3*.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., … Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, *26*(12), 237–245.

Vasu, K., & Nagaraja, V. (2013). Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiology and Molecular Biology Reviews*, *77*(1), 53–72.

The Roles of Possibility and Mechanism in Narrative Explanation

Abstract

There is a fairly longstanding distinction between what are called the *ideographic* as opposed to *nomothetic* sciences. The nomothetic sciences, such as physics, offer explanations in terms of the laws and regular operations of nature. The ideographic sciences, such as natural history (or, more controversially, evolutionary biology), cast explanations in terms of narratives. This paper offers an account of what is involved in offering an explanatory narrative in the historical (ideographic) sciences. I argue that narrative explanations involve two chief components: a possibility space and an explanatory causal mechanism. The presence of a possibility space is a consequence of the fact that the presently available evidence underdetermines the true historical sequence from an epistemic perspective. But the addition of an explanatory causal mechanism gives us a reason to favor one causal history over another; that is, causal mechanisms enhance our epistemic position in the face of widespread underdetermination. This is in contrast to some recent work that has argued against the use of mechanisms in some narrative contexts. Indeed, I argue that an adequate causal mechanism is always involved in narrative explanation, or else we do not have an explanation at all.

1. Introduction

The historical sciences (geology, paleontology, evolutionary biology, etc.)[1] are usually thought to deploy different explanatory strategies than the non-historical sciences (Turner 2007; Turner 2013). Whereas physics, say, seeks explanations given in terms of general laws and the like, the historical sciences seek to explain in terms of narratives. In this paper I will argue for a version of narrative explanation involving two chief components: possibility spaces and causal mechanisms. It has recently been argued that complex historical narratives (to be defined later) can't support explanations involving causal mechanisms (Currie 2014). I argue that this is mistaken. I'll go over some recent work on the history of abiogenesis research to support this contention.

The argument presented in this paper will defend two primary claims: (1) the conceptual structure of narrative explanations nearly always involves a space of alternative possibilities. This can be for either epistemic or ontological reasons. From and epistemic perspective possibility spaces are necessary on account of our position relative to the available evidence. That is, the available evidence radically underdetermines any particular causal history, and on the basis of that fact many possible histories appear compatible with what we know (see Gordon and Olson 1994, p. 15). Construed ontologically, a set of historical facts might involve a high degree of objective contingency—it might be the case that things really could have gone a number of different ways. For the purposes of this paper I remain silent with respect to this ontological aspect and defend the importance of possibility spaces for largely epistemic reasons. (2) Adequate causal mechanisms enhance our epistemic position relative to alternative causal

[1] I note that the idea of evolutionary biology as a properly "historical science" is a controversial one. See Ereshefsky (1992) for some strong arguments against the idea of evolutionary biology as having a distinctively 'historical' flavor.

histories. Causal mechanisms put us in a position to better assess the plausibility of a given history within our possibility space, and in this way enhance the epistemic power of a purportedly explanatory historical narrative. This can involve either the actual discovery of such mechanisms, or raw theoretical innovation. Citing an adequate causal mechanism may not discriminate between possibilities in decisive fashion. Rampant underdetermination seems to rule out such a possibility (see Turner 2007). But an adequate mechanism does make a given explanation more explanatory than its competitors, and so part of the task is to see how this notion of mechanistic adequacy can be cashed out in such a way as to make this notion of *explanatoriness* epistemically significant and not simply *ad hoc*.

2. The Role of Possibility Spaces

In the introduction I said that I would defend two major claims: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. This section will address the first claim by giving a more detailed account of the conceptual structure of narrative explanations and why the role of possibility spaces is so central to them.

When confronted with a natural historical problem (e.g. accounting for the processes involved in the formation of atoll reefs, say (see Ghiselin 1969)) it is my claim that what we are confronted with is, in fact, a space of *possible* histories. That is, when the historical scientist attempts to answer the question, "What geological process accounts for the formation of atoll reefs?" she understands—perhaps implicitly—that there are a number of ways things *might* have gone: she sees many possible histories. This space of possible histories essentially generates a contrasting set of possible explanations, each possible history corresponding roughly to one

hypothetical solution to the problem.[2] Obviously there's just one causal history that actually obtained, but the evidential situation is such that this history is not uniquely fixed from an epistemic perspective (see Roth 2017). The historical scientist's explanatory task then consists in finding the best approximation of the true causal history.

A nice example of this sort of reasoning process can be glimpsed in the debates over speciation processes among evolutionary biologists and paleobiologists. Stephen J. Gould and Niles Eldredge (1972) developed the theory of *punctuated equilibria* to account for the pattern of speciation witnessed in the fossil record. The idea of punctuated equilibria, in brief, holds that evolutionary change occurs in sudden bursts (on geological timescales, anyway), followed by long periods of relative evolutionary stasis. The going theory of evolutionary change at the time held to *phyletic gradualism*—the idea that the pace of evolution is slow and relatively uniform (see Turner 2011). Each of these alternatives is broadly consistent with the available fossil evidence. Phyletic gradualism takes the view that the evolutionary process is gradual, and that the fossil record is very patchy. The putatively patchy character of the fossil record means that we shouldn't expect to be able to use it as a tool for faithfully reading off patterns of speciation in the actual history of life. The theory of punctuated equilibria has it that the fossil record is relatively faithful to evolutionary history, meaning that the fossil record *does* have some explanatory import with respect to uncovering important evolutionary patterns (like speciation). The evidence in the fossil record can support either interpretation.

Consider another example, this time from geology. 19th century geologists were confronted with a fascinating geological puzzle involving what were called 'erratic blocks'.

[2] I'm certainly *not* claiming that the historical scientist is in a position to generate or realize all possible histories, as the number of such alternatives is plausibly infinite. But certainly it's possible to generate quite a few, and it seems that in fact we usually do.

These hulking slabs of (usually) granite are found miles away from any related rocks, and so the obvious question to be answered is, "How did such a large piece of granite come to be deposited here?" In 1820s Europe the answer was not immediately obvious. One well-documented case involved a granite erratic in Switzerland, which was determined to be composed of primary rocks of Alpine origin, but resting on a limestone formation many hundreds of miles from any mountains (see Rudwick 2014, pp. 117-25). Several explanations were offered: that it was deposited by the waters of the Noahic deluge; that it was carried and deposited by waters traveling down the Alps from a broken mountain dam; and only later that it was carried by glacial ice and then deposited after a subsequent melt. The process of adjudicating between each such purportedly explanatory histories (whether evolutionary patterns or seemingly bizarre geological deposits) is the subject of the next section.

It's important to stress that the evidential underdetermination of historical hypotheses is quite different than underdetermination in science more broadly. Turner (2007) argues convincingly that the problem of underdetermination is rather severe in the historical sciences given that natural processes actively destroy the evidential traces on which historical scientists rely.[3] There are two points that make this worthy of note. First, it is precisely for this reason that the explanatory task of the historical scientist *necessarily* involves the generation of a possibility space. If we can think of a natural history as a story concerning the artifacts of the natural world, then what the world presents us with is a story that's missing a great many pages. The unfortunate fact of the matter is that there are many ways of filling those pages in, each of which

[3] Turner appeals to the role played by background theories in the historical sciences to motivate his point. Here, there relevant theory is *taphonomy*, which describes the mechanisms by which the relevant evidence is destroyed (remineralization, decomposition, etc.).

is broadly compatible with our evidential situation.[4] Second, widespread underdetermination is what motivates the earlier insight that the explanatory aspiration of historical science is to give the best *approximation* of the true causal history. It is implausible to think that any of the historical hypotheses we generate will fill in the missing pages perfectly, but we can have reasons to think that some hypotheses outperform others (of which more to come).

To summarize, possibility spaces are ineliminable from narrative explanations because of our epistemic position relative to the evidence at hand. What we want is to develop a causal history that explains the phenomenon in question (e.g. erratic blocks and evolutionary patterns), but right away we realize that many different and mutually incompatible histories could—hypothetically—do the trick. The construction of a space of live possibilities allows us to have some degree of confidence that we've explored the relevant alternatives.[5] Once we've developed a space of possibilities, the initial question (such as, "What accounts for the formation of atoll reefs?") becomes importantly *contrastive*: "Why *x* and not *x'*?" where *x* and *x'* are alternative possible causal histories accounting for the target phenomenon. We want to know how it is that possibilities come to be "foreclosed" upon as a narrative explanation develops, as Beatty (2016) puts it.

3. Causal Mechanisms and Hypothesis Adjudication

[4] See Turner (2011) chapter 2 for more in-depth discussion.

[5] There's a way of reading this that might tempt one to see this as something akin to *inference to the best explanation*. Any such connection is largely superficial. The primary reason for this is that the explanatory scheme that I'm outlining is not meant to be making any especially strong claims about the strength of an explanation as related to its connection to reality. Perhaps none of the causal histories we generate are very accurate as descriptions of the true causal history.

I now turn my attention to an explication and defense of (2): adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. Causal mechanisms are what provide reasons for preferring one possible causal history over another as regards the space of possible histories generated by the natural historical problem at hand.

3.1. Mechanistic set-ups-

Because contingency is generally seen as playing such a fundamental role in natural historical contexts, the relevant mechanisms are not likely to be cashed out in terms of 'invariances' and 'regularities,' as is common in other scientific contexts (see Havstad 2011; Darden and Craver 2002). For the purposes of natural history we might instead think in terms of a more minimal conception of causal mechanisms that I'll call *mechanistic set-ups*. A mechanistic set-up differs from paradigmatic mechanisms (as in Glennan (2002))[6] in that it will often be the case that mechanistic set-ups are the result of one-off circumstances. Paradigmatic mechanisms characterize causal systems that are largely stable across time (think of protein synthesis, for instance). Mechanistic set-ups are not stable across time in this way, but still render outcomes causally expectable given that the right antecedent conditions obtain. That is, given that the right antecedent conditions obtain (and this may, of course, be a *highly contingent* affair), the causal output of the system is fully determined—we have a case of mechanical causal output.

Nancy Cartwright and John Pemberton (2013) give a simple example of a mechanistic set-up using a toy sailboat. When the toy boat is placed in the water it displaces enough liquid to

[6] "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations."

stay afloat; it has a wind-catching device for locomotion; the wind-catching device is acted about by wind gusts in order to achieve locomotive action. If we take this example as having to do with the actions of an *agent* that brings about the mechanistic set-up then we might incline toward an interpretation of the situation in terms of paradigmatic mechanisms. But imagine there's no agent involved at all; that is, let it be the case that nobody placed the boat on the water, and likewise nobody chose any windy day in particular for the use of the boat. Instead suppose that it is a series of contingent events (a child threw the boat in the garbage, it fell out of the garbage truck on the highway, and is now on the surface of a local pond, etc.) that have made things such that the boat is at some later time moving across the top of the water in the expected way.

The one-offness of the circumstances in the revised toy boat example doesn't seem to make the situation non-mechanistic in character. Rather, the mechanism just isn't stable across time in the same way paradigmatic mechanisms are. This is a mechanism in a more minimal sense: it is a mechanistic set-up. In other words, the realization of appropriate antecedent conditions renders the outcome causally expectable, even though the antecedent conditions are highly contingent.[7]

This case is so simple that it won't have much bite against Currie. Recall that Currie's claim is that mechanisms show to be of no use in *complex* narratives. In these cases the explanatory targets are *diffuse*, meaning that they involve complex networks of causal contributors (Currie 2014). An example of a diffuse target is Sauropod gigantism, Gigantism involves, at least, skeletal pneumatization, ovipary, increased basal metabolic rate, etc. Nothing seems to unify such causal contributions, and so there is no *mechanism* for gigantism, according to Currie—the explanatory target is *too diffuse* in complex narratives.

[7] See chapter 3 of Conway Morris (2003) for an in-depth discussion.

3.2. Abiogenesis, mechanistic set-ups, and hypothesis adjudication-

Abiogenesis, I argue, qualifies as a minimal mechanistic set-up in the sense just argued for. That is, the set of facts that determined the development of the very first self-replicating, heterotrophic organisms are plausibly subject to a high degree of contingency (see Conway Morris 2003), but even so, life is a deterministic consequence of just such a contingent set of facts.[8] Further, the instances that the theory aims to explain (e.g. self-replicating molecular systems; heterotrophic metabolic systems; protective membrane enclosures, etc.) are diffuse in the same sense as Sauropod gigantism. My aim here is not to give a full theoretical survey of abiogenesis, but instead to provide just enough content to justify the claim that work in this area fulfills the description of narrative already given, and that causal mechanisms play an important explanatory role, specifically to do with hypothesis adjudication.

Probably the first serious theoretical work on the origins of life is A.I. Oparin's 1923 *The Origins of Life* (Falk and Lazcano 2012). The basic theoretical framework is familiarly Darwinian. Oparin had in mind a model of biological origins whereby life comes on-line in stages, rather than all at once. The prebiotic world, on this view, was one of something approximating 'molecular competition.' For Oparin this amounted to chemical assemblages witnessing differential stability, approximately underwriting a growth model of molecular evolution (Falk and Lazcano 2012; Pigliucci 1999). The primary thing to be explained, on this model, was the development of heterotrophic metabolism. Metabolic pathways are so complex

[8] Some recent work in origins of life research may end up giving reasons to question the assumed contingency of life's emergence. See Kauffman (1993) for a classic treatment of the "self-organization" thesis, and England (2015) for more recent theoretical developments.

that Oparin thought their development must be accounted for in a basically stepwise fashion. Differential stabilities of chemical assemblages would make it such that certain molecules would make up increasingly large proportions of the chemical 'population,' making them live candidates for further downstream innovation (like complex metabolic pathways).

Oparin-type selection models have mostly—though perhaps not entirely—fallen by the wayside. Contemporary work is focused primarily on accounting for the possibility of self-replication and autocatalysis (Penny 2005). The thought is that biological origins must be accounted for in something like a two-step process, one involving the development of self-replicating material suitable for hereditary mechanisms, and another for things like metabolism and heterocatalytic functions like protein construction (Falk and Lazcano 2012; Conway Morris 2003). One of the more promising research strains in this area concerns what's known as the 'RNA World' (Conway Morris 2003). It's widely believed to be the case that the first replicators were RNA (or RNA-like) molecules. So, RNA World researchers are attempting to simulate the conditions of the prebiotic Earth in the laboratory in order to see whether the RNA model of biological origins can carry its empirical weight.

Of note for the purposes of this paper is that the dispute between metabolism-first and replication-first models of abiogenesis is precisely over whether the causal mechanisms in play can adequately account for the target phenomenon: namely, the development of living organisms in the ancient history of Earth. H.J. Muller developed a theoretical agenda stressing the need for self-replicators at the historical foundations of life (Falk and Lazcano 2012). Oparin took heterotrophic metabolic pathways as the primary puzzle to be solved (Oparin 1938; Falk and Lazcano 2012). The replication-first view has emerged as the going view among contemporary researchers primarily because it offers a more plausible mechanism for life's early development.

In order to build complex metabolic pathway it seems like it's first necessary to have a genome space that's large enough to enable downstream innovation of complex functions. So it is that the replication-first view and the research agenda dictated by projects like RNA World are taken to be more explanatory than Oparin-type explanations given in terms of selection among molecular assemblages.

4. Putting Things Together

Let's recall once more the two key claims being advanced: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. Widespread underdetermination in the historical sciences leads to the persistent appearance of possibility spaces as specified by (1), and the development of adequate causal mechanisms specified under (2) enhances our ability to adjudicate the alternatives we're faced with. Causal mechanisms put us in a position to address the contrastive question, "Why *x* and not *x'*?" Causal mechanisms are the devices by which historical counterfactuals become foreclosed upon in the sense of Beatty (2016).

Because explanation in the historical sciences is contrastive in the above sense, I argue that some notion of mechanism is involved in *every* case of successful narrative explanation. Currie (2014) argues that causal mechanisms are appropriate only for the purposes of simple narratives apt to be embedded in terms of regularities. Complex narratives with their diffuse explanatory targets require something more piecemeal that doesn't count as a causal mechanism. My more minimal conception of causal mechanisms given in terms of *mechanistic set-ups* sheds light on why this can't be right. Mechanistic set-ups aren't stable across time like paradigmatic

mechanisms, and yet we have good reason to think that the consequences of such set-ups are mechanistically determined (see Penny 2005; Glennan 2010).[9] It is just this sort of conception of mechanism that helps us to make sense of explanatory success in abiogenesis (such as it is).

Surely the genesis of the first biotic creatures is every bit as diffuse an explanatory target as Sauropod gigantism. I've argued (and I think convincingly) that it is precisely due to the adequacy of some underlying mechanism that one explanatory agenda in abiogenesis has been accepted over the alternatives. The complexity of the narrative and the diffuseness of the explanatory target appear to be beside the point. Without an adequate mechanism—however minimally construed—we can't answer the contrastive question, and so we have no explanation at all.

5. Objection and a Reply

According to Currie (2014) mechanistic set-ups (*ephemeral mechanisms* (Glennan 2010)) look like they're simply pointing to claims about sensitivity to initial conditions. If that's right, then there's a problem, because causal processes in natural historical contexts are often thought to be contingent not just in the sense that they display sensitivity to initial conditions. Such processes are taken to be subject to contingencies in a more robust sense involving "causal cascades" themselves (Currie 2014). It is not unreasonable, for instance, to think that whether a chemical assemblage will manage to hit the right configuration and produce a self-replicating RNA strand is not just a matter of realizing the right set-up conditions (independent of the chances of hitting

[9] Penny notes some interesting experimental results in which living organisms are frozen to near absolute zero, meaning that all information concerning the positions and velocities of the particles in their make-up is lost. They can, nonetheless, be successfully reanimated. Given that the only information that's retained after such a deep freezing involves the chemical structure of the organisms, a natural inference is that 'life' is a mechanical consequence of chemical parts.

on such a configuration). Whether the chemical elements enter into the appropriate causal relations for manifesting autocatalysis might *itself* be a probabilistic matter. Having the right elements might not be all you need—you might need the right elements plus a bit of probabilistic luck. Objective probabilities of this sort might do some damage to the mechanistic account, since it would seem not to be the case that an explanandum *just follows* from a causal set-up. The force of this objection is at least partly dependent on one's answer to the question of where in the world we ought to 'place' objective chances (if there are any).

Most of our intuitions about objective probabilities (probably) derive from our ongoing observations of the world. A lot of stuff in the world *just seems* chancy. We regularly speak in terms of the "odds" or "chances" of developing cancer and the like. Simplifying quite a bit, when we say that there's a 40 percent chance that Susan will live for more than 5 years after being diagnosed with some cancer that has developed to some particular stage, what we're saying is that approximately 40 percent of people that present as cases sufficiently similar to Susan have lived for 5 years or more. One way to read this is in terms of causal indeterminacy. That is, there is really no matter of the fact at time *t* as to what will be the case at time *t'*, aside from the probabilistic facts about cancer populations. The future is (to some degree) causally open, as the causal cascades are operating in a fundamentally probabilistic way.

Such a reading, however, is by no means forced. Bruce Glymour (1998) offers a picture wherein objective probabilities are placed at the level of causal *interactions*. That is, entities *e* and *e** enter into causal interactions with each other on a probabilistic basis, but when they do, the downstream effects unfold in a fully deterministic fashion. Probabilistic partitions of the world, then, are just reflections of whether certain causal interactions became manifest in certain subpopulations or not. If 40 percent of patients with a certain cancer at a particular stage will

survive for more than five year, it's because free radicals (probabilistically) failed to enter into certain causal interactions with healthy cells. The opposite is the case for the contrasting class of fatal cases. On this picture, determinism of the relevant kind seems to be preserved. In such cases as the right causal interactions are realized, downstream effects unfold in mechanical fashion.

6. Conclusion

In this paper I argued for two main claims: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. The reason that narrative explanations involve possibility spaces has to do with our epistemic position relative to the available evidence. Undetermination so permeates the historical sciences that any problem for which we seek an explanation will involve an array of possible alternative causal histories, each of which is broadly consistent with the available evidence. It is the introduction of an adequate causal mechanism that puts us in a position to improve our epistemic lot—with a good mechanism in hand, we can begin to foreclose upon alternatives.

References

Beatty, John. 2016. “What Are Narratives Good For?” *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 58. Elsevier Ltd: 33–40. doi:10.1016/j.shpsc.2015.12.016.

———. 2017. “Narrative Possibility and Narrative Explanation.” *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 1–14. doi:10.1016/j.shpsa.2017.03.001.

Cartwright, Nancy, and John Pemberton. 2013. "Aristotelian Powers: Without Them, What Would Science Do?" in Groff & Greco (Eds.), *Powers and Capacities in Philosophy: The New Aristotelianism*. New York: Routledge.

Conway Morris, Simon. 2003. *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge: Cambridge University Press.

Currie, Adrian Mitchell. 2014. “Narratives, Mechanisms and Progress in Historical Science.” *Synthese* 191 (6): 1163–83. doi:10.1007/s11229-013-0317-x.

Darden, Lindley, and Carl Craver. 2002. "Strategies in the interfield discovery of the mechanism of protein synthesis." *Studies in History and Philosophy of Biological and Biomedical Sciences* 33: 1-28.

Eldredge, Niles, and Stephen J. Gould. 1972. "Punctuated equilibria: an alternative to phyletic gradualism," in Schopf (Ed.), *Models in Paleobiology*. San Francisco: Freeman Cooper.

England, Jeremy. 2015. "Dissipative Adaptation in Self-Driven Assembly." *Nature Nanotechnology*, 10: 919-923.

Ereshefsky, Marc. 1992. "The Historical Nature of Evolutionary Theory." In *History and Evolution*, ed. Matthew Nitecki and Doris Nitecki. New York: The SUNY Press.

Falk, Raphael, and Antonio Lazcano. 2012. “The Forgotten Dispute: A.I. Oparin and H.J. Muller on the Origin of Life.” *History and Philosophy of the Life Sciences* 34 (3): 373–90.

Ghiselin, Michael T. 1969. *The Triumph of the Darwinian Method*. Chicago: Chicago University Press.

Glennan, Stuart. 1996. “Mechanisms and the Nature of Causation.” *Erkenntnis* 44 (1): 49–71.

———. 2002. “Rethinking Mechanistic Explanation.” *Philosophy of Science* 69 (S3): S342–53.

———. 2010. “Ephemeral Mechanisms and Historical Explanation.” *Erkenntnis* 72 (2): 251–66. doi:10.1007/s10670-009-9203-9.

Glymour, Bruce. 1998. “Contrastive, Non-Probabilistic Statistical Explanations.” *Philosophy of Science* 65 (3): 448–71.

Gordon, Malcolm and Everett Olson. 1994. *Invasions of the Land*. New York: Columbia University Press.

Haldane, J.B.S. 1954. "The origin of life." *New Biology* 16: 12-27.

Havstad, Joyce C. 2011. “Problems for Natural Selection as a Mechanism.” *Philosophy of Science* 78 (3): 512–23. doi:10.1086/660734.

Hull, David. 1975. “Central Subjects and Historical Narratives.” *History and Theory* 14 (3): 253–74.

Jeffares, Ben. 2008. “Testing Times: Regularities in the Historical Sciences.” *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 39 (4). Elsevier Ltd: 469–75. doi:10.1016/j.shpsc.2008.09.003.

Kauffman, Stewart. 1993. *The Origins of Order: Self Organization and Selection in Evolution*. Oxford: Oxford University Press.

Mink, Louis O. 1970. "History and Fiction as Modes of Comprehension." *New Literary History*, 1 (3): 541-558.

Oparin, A.I. 1938. *The Origin of Life*. New York: MacMillan.

Penny, David. 2005. “An Interpretive Review of the Origin of Life Research.” *Biology & Philosophy* 20 (4): 633–71. doi:10.1007/s10539-004-7342-6.

Pigliucci, Massimo. 1999. "Where do we come from? A humbling look at the biology of life's origin." *Skeptical Inquirer* 99: 193-206.

Ricoeur, Paul. 1984. *Time and Narrative (Volume 1)*. Chicago: University of Chicago Press.

Roth, Paul A. 2017. “Essentially Narrative Explanations.” *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 1–9. doi:10.1016/j.shpsa.2017.03.008.

Rudwick, M.J.S. 2014. *Earth's Deep History: How It Was Discovered and Why It Matters*. Chicago: Chicago University Press.

Sepkosi, David. 2012. *Rereading the Fossil Record: The Growth of Paleontology as an Evolutionary Discipline*. Chicago: Chicago University Press.

Sunstein, Cass R. 2016. “Historical Explanations Always Involve Counterfactual History.” *Journal of the Philosophy of History* 10 (3): 433–40. doi:10.1163/18722636-12341345.

Turner, Derek. 2013. “Historical Geology: Methodology and Metaphysics.” *Geological Society of America Special Papers* 502 (2): 11–18. doi:10.1130/2013.2502(02).

———. 2007. *Making Prehistory: Historical Science and the Scientific Realism Debate*. Cambridge: Cambridge University Press.

———. 2011. *Paleontology: A Philosophical Introduction*. Cambridge: Cambridge University Press.

A Better Foundation for Public Trust in Science

S. Andrew Schroeder
Claremont McKenna College/Princeton University
aschroeder@cmc.edu

draft of 15 June 2018

Abstract. There is a growing consensus among philosophers of science that core parts of the scientific process involve non-epistemic values. This undermines the traditional foundation for public trust in science. In this paper I consider two proposals for justifying public trust in value-laden science. According to the first, scientists can promote trust by being transparent about their value choices. On the second, trust requires that the values of a scientist align with the values of an individual member of the public. I argue that neither of these proposals work and suggest an alternative that does better: when scientists must appeal to values in the course of their research, they should appeal to *democratic values*, the values of the public or its representatives.

1. Introduction

The American public's trust in science is a complicated matter. Surveys reveal that trust in science has remained consistently high for decades, and scientists remain among the most highly-trusted professional groups (Funk 2017). However, within some segments of society (especially conservatives) trust has declined significantly (Gauchat 2012), and there are obviously serious gaps in trust on certain issues, such as climate change, vaccine safety, and GM foods (Funk 2017). The picture, then, is a complex one, but on balance it is clear that things would be better if the public placed greater trust in science and scientists, at least on certain issues.

As a philosopher, I am not in a position to determine what explains the lack of trust in science, nor to weigh on what will in fact increase trust. Instead, in this paper I will look at the question of what scientists can do to *merit* the public's trust — under what conditions the public *should* trust scientists. Indeed, it seems to me that we need to answer the normative question first: if we take steps to increase

public trust in science, our goal should not simply be to make scientists *trusted*, we should also want them to be *trustworthy*.

In what follows, I'll first explain how recent work in the philosophy of science undermines the traditional justification given to the public for trusting science. I'll then consider two proposals that have been offered to ground public trust in science: one calling for transparency about values, the second calling for an alignment of values. I'll argue that the first proposal backfires — it rationally should *decrease* trust in science — and the second is impractical. I'll then present an alternative that is imperfect, but better than the alternatives: when scientists must appeal to values in the course of their work, they should appeal to *democratic values* — roughly, the values of the public or its representatives.

2. Trust and the Value-Free Ideal

Why should the public trust scientists? The typical answer to that question points to the nature of science. Science, it is said, is about facts, and not values. It delivers us objective, verifiable truths about the world — truths not colored by political beliefs, personal values, or wishful thinking. Of course, there are scientists who inadvertently or intentionally allow ideology to influence their results. But these are instances of *bad* science. Just as we should not allow the existence of incompetent or corrupt carpenters to undermine our trust in carpentry, we should not allow the existence of incompetent or corrupt scientists to undermine our trust in science. So long as we have institutions in place to credential good scientists and root out corrupt ones, we should trust the conclusions of science.

There is, unfortunately, one problem with this story: science isn't actually like that. In the past few decades, philosophers of science have shown that even good science requires non-epistemic value judgments. Without wading into the nuanced differences between views, I think it is fair to say that there is a consensus among philosophers of science that non-epistemic values can appropriately play a role in at

least some of the following choices: selecting scientific models, evaluating evidence, structuring quantitative measures, defining concepts, and preparing information for presentation to non-experts.[1]

These value choices can have a significant impact on the outcome of scientific studies. Consider, for example, the influential Global Burden of Disease Study (GBD). In its first major release it described itself as aiming to "decouple epidemiological assessment from advocacy" (Murray and Lopez 1996, 247). In the summary of their ten volume report, the authors describe their study as making "a number of startling individual observations" about global health, the first of which was that, "[t]he burdens of mental illnesses…have been seriously underestimated by traditional approaches… [P]sychiatric conditions are responsible…for almost 11 per cent of disease burden worldwide" (Murray and Lopez 1996, 3). Many others have cited and relied on the GBD's conclusions concerning the magnitude of mental illness globally (Prince *et al*. 2007). And nearly two decades later, the same GBD authors, in commenting on the legacy of the 1996 study, proudly noted that it "brought global, regional, and local attention to the burden of mental health" (Murray *et al*. 2012, 3).

It turns out, however, that the reported burden of mental health was driven largely by two value choices: the choice to "discount" and to "age-weight" the health losses measured by the study. Discounting is the standard economic practice of counting benefits farther in the future as being of lesser value compared to otherwise similar benefits in the present, and age-weighting involves giving health losses in the middle years of life greater weight than otherwise similar health losses among infants or the elderly. Further details about discounting and age-weighting aren't relevant to this paper; all we need to note is that the study authors acknowledged that each reflects value judgments, and that a reasonable case could be made to omit them (Murray 1996; Murray *et al*. 2012).[2] Given other methodological choices made by the authors, these two weighting functions combine to give relatively more weight to health

[1] On these points see e.g. Reiss (2017) and Elliott (2011).

[2] Indeed, in 2012 the GBD ceased age-weighting and discounting. There was also a third value choice that drove the large burden attributed to mental health: the choice to attribute all suicides to depression (Murray and Lopez 1996, 250). Because I do not know precisely how this affected the results, I set it aside here. For much more on discounting, age-weighting, and other value choices in the GBD, see Schroeder (2017).

conditions which (1) commonly affect adults or older children (rather than the elderly or young children), (2) have disability (rather than death) as their primary impact, and (3) have their negative effects relatively close to the onset or diagnosis of the condition (rather than far in the future). It should not be surprising, then, that when the GBD authors ran a sensitivity analysis to see how the decision to discount and age-weight affected the results, they discovered that the conditions most affected by these choices — unipolar major depression, anaemia, alcohol use, bipolar disorder, obsessive-compulsive disorder, chlamydia, drug use, panic disorder, post-traumatic stress disorder — were largely composed of mental health conditions (Murray and Lopez 1996, 282). Overall, the global burden of disease attributable to psychiatric conditions drops from 10.5% to 5.6%, when the results are not age-weighted or discounted (Murray and Lopez 1996, 261, 281).

I don't want to comment here on the wisdom of the GBD scientists' decision to discount and age-weight.[3] They offer clear arguments in favor of doing so and many other studies have done the same, so at minimum I think their choices were defensible. The point is that what was arguably the top-billed result of a major study — a result which was picked up on by many others, and which was still being proudly touted by the study authors years later — was not directly implied by the underlying facts. It was driven by a pair of value judgments. Had the GBD scientists had different views on the values connected to discounting and age-weighting, they would have reported very different conclusions concerning the global impact of mental illness.[4]

This case is not unique. The dramatically different assessments given by Stern and Nordhaus on the urgency of acting to address climate change can largely be traced to the way each valued the present versus the future (Weisbach and Sunstein 2009). Similar conclusions are plausible concerning the value choices involved in classifying instances of sexual misbehavior in research on sexual assault, the value

[3] I do so in Schroeder (unpublished-a).

[4] Although the sensitivity analysis was conducted by the original study authors, they do not draw any connection to their prominent claims concerning the global extent of mental illness. To my knowledge, this paper is the first to do so.

choices impacting the modeling of low-level exposures to toxins (Elliott 2011), and the value choices involved in constructing price indices (Reiss 2008).

A natural — and not implausible — response to these cases is to suggest they are outliers. Although some scientific conclusions are sensitive to value choices, the vast majority are not. The Earth really is getting warmer and sea levels really are rising, due to human activity. Vaccines really do prevent measles and really don't cause autism. These conclusions are not sensitive in any reasonable way to non-epistemic value judgments made by scientists in the course of their research. The problem, however, is that there is no clear way for a non-expert to verify this — to tell which cases are the outliers and which are not. This, I think, justifies a certain amount of skepticism. "Although some of our conclusions do depend on value judgments, trust us that *this* one doesn't," isn't nearly as confidence-inspiring as, "Our conclusions depend only on facts, not values."

I conclude, then, that rejecting the view of science as value-free, combined with high-profile examples of scientific conclusions that do crucially depend on value judgments, undermines the claim of science to public trust in a significant way. In other words, it explains why it may be rational for the public to place less trust in the conclusions of science on a broad range of issues — including in areas, such as climate change and vaccine safety, where major conclusions are not in fact sensitive to different value judgments.[5]

3. Grounding Trust in Transparency

Good science is not value-free, which undermines the standard justification given for trust in science. What, then, can scientists do to merit the public's trust? The standard response has been to appeal to transparency. If values cannot or should not be eliminated from the scientific process, scientists

[5] For similar conclusions see Douglas (2017); Wilholt (2013); Irzik and Kurtulmus (*forthcoming*); and Elliott and Resnik (2014).

should be "as transparent as possible about the ways in which interests and values may influence their work" (Elliott and Resnik 2014, 649; *cf.* Ashford 1998; Douglas 2008; McKaughan and Elliott 2018). Obviously, in order for this proposal to work, scientists would need to be aware — much more aware than most are today — of the ways in which value judgments influence their work. But, since we have independent reason to want such awareness, let us assume that calls for transparency are accompanied by a mechanism for increasing such awareness by scientists.

Would such a proposal work? Transparency about values can help ground trust in some situations, but I see no reason to think that it should broadly support public trust in science. Transparency is only useful in supporting — as opposed to eroding — trust if it enables the recipient of that information to determine how it has affected the author's conclusions. (Knowing I have a conflict of interest will typically reduce your trust in what I tell you, unless you can determine how that conflict influenced my conclusions.) Transparency, then, will only promote trust in a robust way if the public understands how value choice influenced the results, and understands what alternative value choices could have been made and how they would have influenced the results. These criteria may be satisfiable when the effect of a value choice is relatively simple. Suppose, for example, that a scientist classifies non-consensual kissing as "sexual assault", rather than "sexual misconduct", on the grounds that she believes it has more in common with rape (a clear instance of sexual assault) than it does with contributing to a sexualized workplace (a clear instance of sexual misconduct). The value judgment here is relatively simple to explain, an alternative classification is obvious, and (if the statistics involved are simple) the effect of alternative classification on the study may be relatively straightforward. So transparency could work here.

Many value choices, however, are much more complex. Think about choices embedded in complex statistical calculations — for example, those involved in aggregating climate models (Winsberg 2012) or in calculating price indices (Reiss 2008). In cases like these, it will be very hard to clearly explain the importance of any individual value choice and harder still to explain what alternative choices

could have been made. Further, many studies involve a large number of value judgments. Schroeder (2017), for example, identifies more than ten value choices which non-trivially influenced the Global Burden of Disease Study's results. Even if each of those value choices could be explained individually, it would be virtually impossible for a non-expert to figure out the interaction effects between them.

What these cases show is that even if scientists make a serious effort at transparency — not simply listing their value judgments, but attempting to explain how those judgments have influenced their results — in many cases it simply won't be possible to communicate to the public how those values have impacted their work.[6] And, if the public can't trace the impact of those values, transparency doesn't amount to much more than a warning — a reason to *distrust*, rather than to trust. A parallel realization can be seen in the way many medical schools and journals have handled researchers' conflicts of interest. Whereas in the past disclosures of conflicts of interest — essentially, transparency — were often regarded as sufficient; many have now realized that merely knowing about such conflicts does not appreciably help a reader to interpret a study. There is thus a growing move towards banning all significant conflicts of interest.[7]

4. Grounding Trust in an Alignment of Values

The previous section argued that transparency about values is not typically a solution to the problem of public trust in science. That problem, we can now see, was not caused by the fact that values were *hidden*; it was caused by the fact that the values of scientists may *diverge* from the values of any

[6] McKaughan and Elliott (2018, and in other works) suggest that scientists, through a particular sort of transparency, seek to promote "backtracking" — that is, to enable non-experts to understand how values have influenced scientists' results and to see how those results might have looked given alternative values. They seem to suggest that, at least in the cases they consider, this will frequently be possible. I am claiming that this will not generally be feasible. See Schroeder (unpublished-a) for a more detailed discussion of a particular case.

[7] See e.g. <https://ari.hms.harvard.edu/interim-policy-statement-conflicts-interest-and-commitment>

individual member of the public.[8] To promote public trust in science, then, it seems that we need to eliminate that divergence. This is the insight that motivates Irzik and Kurtulmus (*forthcoming*; *cf*. Douglas 2017; Wilholt 2013), who argue that what they call "enhanced" trust requires that a member of the public knows that a scientist has worked from value choices that are in line with her own.

If this proposal were feasible, I think it would provide a good foundation for trust. And, in certain limited cases, it may be feasible. When science is conducted by explicitly ideological organizations, members of the public may be able to make quick and generally accurate judgments about what values scientists hold, and accordingly may be able to seek out research done by scientists who share their values. (A pragmatic environmentalist, for example, might be confident that scientists employed by the Environmental Defense Fund are likely to share her values.)

Most science, however, is not conducted by explicitly ideological organizations. In these cases, it will typically be very hard for members of the public to confidently determine whether a given study relied on value judgments similar to her own. Even when this can be done (perhaps as a result of admirable transparency and clarity on the part of a scientist), it will require sustained and detailed engagement from the public, who will have to pay close attention not just to the conclusions of scientific studies, but also to their methodology. Although such close attention to the details of science would be beneficial for a great many reasons, it unfortunately is not realistic on a broad scale. There are simply too many scientific studies out there that are potentially relevant to an individual's decisions for even attentive members of the public to keep up. If our model for trust in science requires an alignment of values between the scientist and individual members of the public, trust in science can't be a broad phenomenon. Further, I don't think we want our foundation for trust in science to make that trust accessible only to those with the education and time to invest in exploring the details of individual scientific studies.

[8] It seems relevant to note here that distrust in science is greatest among those who identify as politically conservative, while studies show that university scientists in the U.S. overwhelming support liberal candidates for political office. Whether or not this in fact explains the distrust conservatives have in science, the argument thus far shows why such distrust could have a rational foundation.

I also — somewhat speculatively — worry that adopting this proposal would exacerbate another problem. Suppose the proposal works and, at least on some issues, members of the public are able to identify and rely upon science conducted in accordance with their own values. This, I think, might lead to a further "politicization" of science, as each side on some issue seeks scientists who share their values. Of course, once we allow a role for values in science, value-based scientific disagreement isn't necessarily a problem. Faced, for example, with one experimental design that is more prone to false positives and another that is more prone to false negatives, either choice may be scientifically legitimate. It may therefore be appropriate for more environmentally-minded citizens to rely on different studies than citizens more concerned about economic development. I worry, though, that in a culture where the public specifically seeks science done by those who share their values, it will be too easy to write off any differences in conclusions as due to value judgments — too easy for environmentalists to assume that any time pro-environment and pro-industry scientists reach different conclusions, it must be due to different underlying, legitimate value judgments. In reality, though, most such disagreements are the result of *bad* science. The worry, then, is that if we grow too comfortable with each side of an issue having its own science, it will be harder to distinguish scientific disagreements that can be traced to legitimate value judgments, from disagreements that are based on illegitimate value judgments or simple scientific error. This would be a major loss.

5. Grounding Trust in Democratic Values

I've argued that neither transparency about values nor an alignment of values can provide a broad foundation for public trust in science. Let me, then, suggest a proposal that, though imperfect, can do better. From what's been said so far, we can note a few features that a better solution should have. First, both the transparency and aligned values proposals ran into trouble because they require a great deal of attention and sophistication from the public. Most individuals simply don't have the training to

understand more technical value choices, or value choices embedded within complex calculations. And, even when such understanding is possible, it will often require a level of attention that will in practice be accessible only to the well-off. We should therefore look for a foundation for public trust which doesn't require such detailed understanding of or close attention to individual scientific studies. Second, I suggested that the aligned values proposal, in telling individuals to seek out studies conducted in accordance with their own values, could reinforce a kind of politicization that may have bad consequences. It would be better to find a proposal that wouldn't so easily divide scientists and the public along ideological lines. Third, the problem with the transparency proposal (which the aligned values proposal tried, impractically, to address) was that values, even if transparent, can be alien. In order for an individual to truly trust science, that science must be built on values that have some kind of legitimacy for her.

I think scientists can satisfy two-and-a-half of these three criteria by appealing to *democratic values* — the values of the public and its representatives — when value judgments are called for in the scientific process. The details of this proposal go beyond what I can say here.[9] But, briefly, the idea is that we look to political philosophy to tell us how to determine the (legitimate) values representative of some population. In some cases, those values might be the output of a procedure, such as a deliberative democracy exercise, a citizen science initiative, or a public referendum.[10] In other cases, it might be more appropriate to equate a population's values with the views, suitably "filtered" and "laundered", currently held by its members. ("Filtering" may be necessary to remove politically illegitimate values, e.g. racist values, and "laundering" to clean up values that are unrefined or based on false empirical beliefs.) In cases where there is a broad social consensus, that might count as the relevant democratic value; in cases where there is a bimodal distribution of values, we might say that there are two democratic values; etc.

---

[9] See Schroeder (unpublished-b) for a bit more. Many other philosophers have argued that there should be an important place for democratic values in science. See, for example, Kitcher (2011), Intemann (2015), and Douglas (2005).

[10] The extensive literature on "mini-publics" offers a promising starting point. See e.g. Escobar and Elstub (2017).

Suppose, then, that political philosophers, informed by empirical research, can give us a way of determining democratic values. I suggest that when value judgments are called for within the scientific process,[11] scientists should use democratic values when arriving at their primary or top-line results — the sort of results reported in an abstract, executive summary, or in the initial portions of the analysis. Scientists could then offer a clearly-designated alternative analysis based on another set of values, e.g. their own. I think this proposal can address two of the concerns with which I began this section, and can make some progress towards answering the third.

Let us first consider the too-much-attention and politicization problems. On the democratic values proposal, if an individual can trust that a study was competently carried out — a matter I'll return to below — then she can know, without digging into the methodological details, that its conclusions are based on objective facts plus democratic values.[12] This means that, in most cases, the public need not pay detailed attention to the methodological details of individual studies — thus solving the too-much-attention problem. Further, if scientific conclusions are based on objective facts plus democratic values, any two scientists investigating the same problem in the same social and political context should reach roughly the same conclusion. This recovers a kind of objectivity for science — not objectivity as freedom from values, but objectivity as freedom from personal biases. On this picture, the individual characteristics of a scientist should have no impact on her conclusions — a conception of objectivity that has been defended on independent grounds (Reiss and Sprenger 2014; *cf.* Daston and Galison 2007 on "mechanical objectivity"). If they are both doing good science, the environmentalist and the industrialist should reach the same top-line conclusions. And if the environmentalist and industrialist reach different

[11] This proposal is restricted to value judgments that arise within the scientific process. In particular, I do not mean for it to apply to problem selection. Scientists should be free to choose research projects that are not the projects that would be chosen by the general public. (The public, however, is under no obligation to fund such projects.) I treat the choice of research topics differently than choices that arise within the course of research because I think that scientists have different rights at stake in each case. For some related ideas, see Schroeder (2017b).

[12] There may also, of course, be methodological choices not based on non-epistemic values (including choices based on epistemic values). I set these aside here, since the problems of trust I'm concerned with don't arise in the same way from them.

top-line conclusions, it means that one or the other has made some sort of error. This, I think, provides a solution to the politicization problem: on the democratic values proposal, good science (at least in its primary analyses) will speak with a single voice.

The democratic values proposal therefore solves two of the three problems we noted above. Of course, it only does so if the public can be confident that scientists really are making use of democratic values. Why should the public assume that? Right now, I think the answer is: they shouldn't! For the democratic values proposal to work, it must be accepted by a significant portion of the scientific community, or by an easily-identifiable subset of the scientific community. If that were to happen, though, then the problem here becomes the more general one of how the public can trust scientists to enforce their own norms. The procedures and policies now in place work reasonably well, I think, to expose unethical treatment of research subjects, falsification of data, and certain other types of misconduct. I am therefore optimistic that, given a greater awareness of the role value judgments play in scientific research, a system could be devised to identify scientists who depart from a professional norm requiring the use of democratic values.

6. Science, Values, and Democracy

I've argued that the democratic values proposal can address two of the problems that faced the alternative views. But what about the third? On the transparency proposal, the values of scientists can truly be alien. If a scientist conducts research based on her own values, then, unless I happen to share those values, I have no meaningful relationship to those values. If, however, a scientists appeals to democratic values, then there is a relationship, even if I don't share those values. If democratic procedures or methods were carried out properly, then my values were an input into the process which yielded democratic values. My values are, in a sense, represented in the output of that process. This, in turn, means that those values should have a kind of legitimacy for me. In a democracy, we regularly

impose non-preferred outcomes on people when they are out-voted. So long as democratic procedures are carried out properly, this seems to be legitimate — not ideal, perhaps, but better than any available alternative. On the democratic values proposal, then, when a particular scientific conclusion is uncontested, the public can trust that that conclusion is one drawn solely from the facts, plus perhaps the values that *we* share. For most of us, who don't have the time, inclination, or ability to dig into the details of each scientific study we rely on, or who have a strong commitment to democracy, that will be enough.

I think that the foregoing provides a reasonable answer to the alien values concern. It is of course not a perfect answer. It would be better, at least from the perspective of trust, to get each member of the public access to "personalized" science conducted in accordance with her values. This, however, is impractical, as we saw when discussing the aligned values proposal. So long as that is the case, there is no way to accommodate everyone. Democratic values seem like a reasonable compromise in such a situation.

All of that said, it would be nice if we could say a bit more to those ill-served by democratic values. What should we say, for example, to an individual who knows that her values lie outside the political mainstream on some issue and is therefore distrustful of science done with democratic values on that issue? The first thing to note is that, in such cases, the democratic values proposal fares no worse (or at least not much worse) than the transparency or aligned values proposals. The democratic values proposal is fully consistent with transparency - something we have independent reason to want. So, in cases where the transparency proposal works (e.g. cases where the value choices are few, easy to understand, and computationally simple), the same advantages can be had with the democratic values proposal. Individuals who disagree with a particular value judgment and have the time and expertise to do so can determine how results would have looked under a different set of value judgments. Also, recall that I am proposing only that primary or top-line results be based on democratic values. In cases where value judgments can make a big difference — as in the Global Burden of Disease Study case discussed earlier — we might hope that scientists who hold contrary values will note the dependence of those

results on values by offering secondary, alternative analyses that begin from different value judgments. Those who have the time and expertise to dig into the methodology of scientific reports can do so, seeking out results based on values they share, as the aligned values proposal would recommend.

If the foregoing is correct, the democratic values proposal does better than the alternatives in most cases, and no worse in others. That should be sufficient reason to prefer it. But I think we can say a bit more. In what cases is the complaint from minority values most compelling? It is not, I think, when it comes from people whose values lie outside the mainstream on some issues, but within the mainstream on many other issues. The much more compelling complaint comes from people whose values consistently lie outside the mainstream — people who are consistently out-voted. Oftentimes (though of course not always) when this happens, it involves individuals who are members of groups that are or have been marginalized by mainstream society. Think, for example, of cultural or (dis)ability-based groups whose values and ways of life have been consistently treated as being less valuable and worthy of respect than the values and ways of life of the majority.

I think the democratic values proposal has two important features that can partially address such complaints. First, remember that the democratic values proposal launders and filters the actual values held by the public. Certain values — e.g. racist or sexist ones — conflict with basic democratic principles of equal worth, and so cannot be candidate democratic values. Thus, even in a racist society, telling scientists to work from democratic values will not tell them to work from racist values.[13] Second, in what I regard as its most plausible forms, democracy is not a form of government based on one person-one vote. It is a form of government based on the idea that all citizens are of equal worth and have a right to equal consideration. This suggests that, in cases where minority values are held by a group that is or has been the subject of exclusion or discrimination, democratic principles may sometimes require giving those values extra weight, or a voice disproportionate to their statistical representation in the population, as a way of accounting or compensating for past unjust treatment. Thus, democratic principles may in

[13] See Schroeder (unpublished-b) for more on this.

some cases require treating the values held by an excluded minority as democratically on a par with the conflicting values held by the majority.[14]

These considerations, I think, lessen the force of the complaint from minority values, especially in its most serious incarnation. But I don't think they eliminate it. There will still be people whose values will consistently be marginalized by the democratic view. In such cases, the main recourse available is an appeal to alternate results. If individuals with minority views can count on there being scientists who share those views, they can expect that the kind of alternative analysis they would prefer will be out there, at least in cases where it makes a difference. Of course, scientists are currently a rather homogeneous bunch along many dimensions. So this suggests that the call to work from democratic values provides (yet further) support for the importance of increasing diversity within the scientific community.[15]

[14] See Kelman (2000) for an example of this sort of argument in the context of disability.

[15] ACKNOWLEDGEMENTS TO BE ADDED

References

Ashford, Nicholas. 1988. "Science and Values in the Regulatory Process." *Statistical Science* 3.

Daston, Lorraine and Peter Galison. 2007. *Objectivity*. MIT Press.

Douglas, Heather. 2017. "Science, Values, and Citizens." In *Eppur si muove: Doing History and Philosophy of Science with Peter Machamer,* ed. Adams, Biener, Feest, and Sullivan. Dordrecht: Springer.

—. 2008. "The Role of Values in Expert Reasoning." *Public Affairs Quarterly* 22.

—. 2005. "Inserting the Public into Science." In *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*, ed. Maasen and Weingart. Dordrecht: Springer.

Elliott, Kevin. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. Oxford: Oxford University Press.

Elliott, Kevin and David Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives* 122.

Escobar, Oliver and Stephen Elstub. 2017. "Forms of Mini-Publics: an Introduction to Deliberative Innovations in Democratic Practice," NewDemocracy Research and Development Note, available at <https://www.newdemocracy.com.au/research/research-notes/399-forms-of-mini-publics>.

Funk, Cary. 2017. "Real Numbers: Mixed Messages about Public Trust in Science." *Issues in Science and Technology* 34.

Gauchat, Gordon. 2012. "Politicization of Science in the Public Sphere: A Study of Public Trust in the United States, 1974 to 2010." *American Sociological Review* 77.

Intemann, Kristin. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5.

Irzik, Gürol and Faik Kurtulmus. *Forthcoming*. "What is Epistemic Public Trust in Science?" *British Journal for Philosophy of Science*.

Kelman, Mark. 2000. "Does Disability Status Matter?" In *Americans with Disabilities: Exploring Implications of the Law for Individuals and Institutions,* eds. Francis and Silvers. Routledge

Kitcher, Philip. 2011. *Science in a Democratic Society*. Amherst, NY: Prometheus.

McKaughan, Daniel and Kevin Elliott. 2018. "Just the Facts or Expert Opinion? The Backtracking Approach to Socially Responsible Science Communication," in *Ethics and Practice in Science Communication* (eds. Priest, Goodwin, and Dahlstrom). Chicago: University of Chicago Press.

Murray, Christopher. 1996. "Rethinking DALYs." In *The Global Burden of Disease,* ed. Murray and Lopez.

Murray, Christopher and Alan Lopez (Eds). 1996. *The Global Burden of Disease*. Harvard University Press.

Murray, Christopher *et al*. 2012. Supplementary appendix to "GBD 2010: design, definitions, and metrics." *Lancet* 380.

Prince, Martin *et al*. 2007. "No health without mental health." *Lancet* 370.

Reiss, Julian. 2017. "Fact-value entanglement in positive economics." *Journal of Economic Methodology* 24.

—. 2008. *Error in Economics: The Methodology of Evidence-Based Economics*. London: Routledge.

Reiss, Julian and Jan Sprenger. 2014. "Scientific Objectivity." In *The Stanford Encyclopedia of Philosophy* (Winter 2017 edition), ed. Zalta.

Schroeder, S. Andrew. 2017. "Value Choices in Summary Measures of Population Health." *Public Health Ethics* 10.

—. 2017b. "Using Democratic Values in Science: an Objection and (Partial) Response," *Philosophy of Science* 84.

—. Unpublished-a. "Which Values Should We Build Into Economic Measures?" *Under review.*

—. Unpublished-b. "Communicating Scientific Results to Policy-makers," *manuscript on file with author.*

Weisbach, David and Cass Sunstein. 2009. "Climate Change and Discounting the Future: A Guide for the Perplexed," *Yale Law and Policy Review* 27.

Wilholt, Torsten. 2013. "Epistemic Trust in Science." *British Journal for Philosophy of Science* 64.

Winsberg, Eric. 2012. "Values and Uncertainties in the Predictions of Global Climate Models." *Kennedy Institute of Ethics Journal* 22.

PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association

Inferential power, formalisms, and scientific models

Ardourel Vincent*, Anouk Barberousse†, Cyrille Imbert§

* IHPST — CNRS, Université Paris 1 Panthéon-Sorbonne

† SND — CNRS, Sorbonne Université

§ Archives Poincaré — CNRS, Université de Lorraine

Abstract

Scientific models need to be investigated if they are to provide valuable information about the systems they represent. Surprisingly, the epistemological question of what enables this investigation has hardly been investigated. Even authors who consider the inferential role of models as central, like Hughes (1997) or Bueno and Colyvan (2011), content themselves with claiming that models contain *mathematical resources* that provide *inferential power*. We claim that these notions require further analysis and argue that mathematical formalisms contribute to this inferential role. We characterize formalisms, illustrate how they extend our mathematical resources, and highlight how distinct formalisms offer various inferential affordances.

1. Introduction. When analyzing scientific representations, philosophers of science are keen on mentioning that some models provide scientists with "mathematical resources" and "inferential power", but they seldom give a detailed analysis of these notions. This paper is devoted to the discussion of what appears to us as major mathematical resources, namely, formalisms. We thus present an analysis of the notion of formalism as well as examples from which we argue that formalisms should be acknowledged as major units of scientific activity.

We proceed as follows. In Section 2, we briefly review what philosophers of science have to say about mathematical resource and inferential power and observe that it is disappointing. In order to fill the gap we have identified, we put forward in Section 3 the three components we identify within the notion of mathematical resource. Section 4 is devoted to one of these components, namely, formalism. At last, in Section 5, we provide the reader with examples of how the choice of a formalism influences the type of knowledge scientists may draw from their representations.

2. Scientific representations and inferences therefrom. At what conditions can scientific models be used to gain information about target systems? First, a suitable semantic relation between the model and the system(s) that it stands for should obtain, so that by investigating the model, we can make legitimate inferences about its target system(s). This cannot be done unless nontrivial inferences about the model itself, as a mathematical object, can be carried out. Models are usually referred to by proper names (like "Ising model" or "Lotka-Volterra" model") or by expressions that highlight some of their mathematical properties (like "the harmonic oscillator" or "the ideal gas"). There is however more to be learnt about them than their prima facie properties. For example, solving the Ising model reveals more about Ising-like systems than their description as "sets of discrete variables representing magnetic dipole moments of atomic spins that can be in one of two states"; similarly, the mathematical content of an harmonic oscillator goes beyond "being a system that, when displaced from its equilibrium position, experiences a restoring force that is proportional to the displacement". Philosophers of science are aware of the need to investigate the epistemology of models and how we find out about concealed truths about model systems (Frigg,

2010, 257) but are surprisingly silent about how it is actually performed.[1] They are content with saying that the model is "manipulated" (Morgan and Morrison, 1997, chapter 2, passim) or that we can "play" with it (Hughes, 2010, 49), which are suggestive, but metaphoric characterizations.

Surprisingly, even accounts of applied mathematics and scientific representation that give central stage to their inferential role hardly analyze how it is fulfilled and which elements of the models contribute to it. Let us illustrate this point with Bueno's and Colyvan's work. They claim that "the fundamental role of applied mathematics is inferential" (Bueno and Colyvan, 2011, 352) and accordingly propose an "inferential conception" of the application of mathematics that extends Hughes' three-step DDI account of scientific representation (see below).[2] First, a "mapping from the empirical set up to a convenient mathematical structure" (ibidem, 353) is established (immersion step); by doing so, it becomes possible "to obtain inferences that would otherwise be extraordinarily hard (if not impossible) to obtain" (ibidem, 352) (derivation step); finally, the mathematical consequences that were obtained are interpreted step in terms of the initial empirical set up (ibidem, 353) (interpretation step). Bueno and Colyvan further highlight the importance of the inferential role of mathematics for mathematical unification, novel predictions by mathematical reasoning or mathematical explanations (ibidem, 363). However, the analysis of how this inferential role is carried out shines by its absence. Bueno and Colyvan mostly analyze mathematical resources in a semantic perspective[3] and insist on the difference in content and interpretation that these make possible, e.g., when "mathematics provides additional entities to quantify

[1] Frigg, while clearly stating the problem, does not really address it and is content with briefly emphasizing the advantages of his fictional account of model concerning the epistemology of models (Frigg, 2010). As to the epistemological section of Frigg and Hartmann's review article about scientific models, it merely points at experiments, simulations, thought-experiment as ways of investigating models (Frigg and Hartmann, 2017).

[2] Suarez's inferential conception (Suarez, 2004) hardly addresses either the question of how inferences from models are actually carried out. For lack of space, we shall not discuss it here.

[3] Their discussion is mostly directed at the shortcomings of Pincock's "mapping account" of the application of mathematics (Pincock, 2004).

over" (complex numbers), or is "the source of interpretations that are physically meaningful" and provide "novel prediction" about physical systems, like with the case of the interpretation of negative energy solutions to Dirac's equation (ibidem, 366).

In another paper, Bueno suggests that results are derived "by exploring the mathematical resources of the model" in which features of the empirical set up are immersed (Bueno, 2014, 379, see also 387) and that results emerge "as a feature of the mathematics" (ibidem) or by using "the particular mathematical framework" (ibidem, 385). What this inferential power of mathematics should be specifically ascribed to remains unclear. Bueno and Colyvan (2011, 352) just claim that the "embedding *into a mathematical structure* makes it is possible to obtain inferences". They also emphasize how, with the help of appropriate idealizations, "the *mathematical model* [can] directly [yield] the results" (ibidem, 360, our emphasis). But elsewhere in the paper, consequences are said to be drawn "*from the mathematical formalism*, using the mathematical structure obtained in the immersion step" (ibidem, 353, our emphasis).

What are we to make of these various claims? A prima facie plausible answer to this question might be that structures and formalisms are the two sides of a same inferential coin. However, this answer is not satisfactory, since, as is well-known, mathematical structures can be presented in different formalisms, which, as we shall see in Section 4, are associated with different inferential possibilities. Another blind spot in Bueno's and Colyvan's account is that while the derivation step is claimed to be "the *key point* of the application process, where consequences from the mathematical *formalism* are generated'' (ibidem, 353), the question of how inferences are drawn with the help of formalisms is left under-discussed.

We draw from this brief analysis of Bueno's and Colyvan's views that the notions of mathematical resource and inferential power, which are commonly used when discussing applications of mathematics, are often mere labels in need of further investigation. Coming back to the seminal ideas presented by Hughes and extended by Bueno and Colyvan is of little help because Hughes' paper lacks precise answers to the following precise questions: What are exactly mathematical resources? What is their inferential power? In his DDI (Denotation, Demonstration, and Interpretation) account of scientific representation, Hughes claims that scientific representations have an ''internal dynamic'', whose effects we can examine (1997, 332), and "contain *resources* which enable us to demonstrate the results we are interested in". A general notion of resource is appropriate to capture the variety of ways in which demonstrations can be

carried out; however, the claim that the deductive power comes from "the *deductive resources* of mathematics they employ" (ibidem, 332) is too vague and is left unanalyzed.

3. Components of mathematical resources. How are the notions of inferential power and mathematical resources to be analyzed? Are they linked to structures or to symbolic systems and formalisms? In this section, we claim that formalisms are an important component of the notions of inferential power and mathematical resource and should be analyzed in their own right.

Let us begin by briefly presenting what are, according to us, the three main components of the notions of mathematical resource and associated inferential power. First, mathematical structures, *to the extent that they are tractable*, are undoubtedly an important part of the mathematical resources that are used in mathematical modeling. As argued by Cartwright, theories are no "vending machines" that "drop out the sought-for representation" (1999, 247); scientific models are no vending machines either and scientists must make the best of the models that they know to be tractable. Accordingly, the content of models often needs to be adapted by means of idealizations, approximations (Redhead 1980), abstractions, by squeezing representations into the straightjacket of a few elementary models (Cartwright, 1981), or by drawing, from the start, on the pool of existing tractable models (Humphreys, 2004, Barberousse and Imbert, 2014).

Second, mathematical knowledge associated with structures is also to be counted as a distinct mathematical resource, which allows for new inferences when it is available. Let us take the well-known example of Koenigsberg's seven bridges. The impossibility of crossing them once and only once in a single trip can be demonstrated by applying a result from graph theory. Similarly, the explanation of the life-cycle of the Magicicada (Baker 2009, Colyvan 2018) is provided by the application of a number-theoretic property of prime numbers to life-cycles of species.

At last, formal settings or formalisms provide languages in which theories are developed, calculations carried out, and inferences drawn from models. Examples of formalisms are Hamiltonian formalism, path integrals, Fourier representation, cellular automata, etc. We provide a detailed analysis of some of these below. Contrary to mathematical structures, formalisms are partly content neutral (though form and content are often intertwined in scientific representations). As providing a partially stan-

dardized way of making inferences, they are important tools for scientists, which in turn justifies considering them as important units of analysis in the philosophy of science. Other authors have started exploring the idea that format matters in scientific activities. Humphreys gives general arguments to this effect and emphasizes the difference between formats that are appropriate for human-made and format that suit computational inferences (2004). Vorms (2009) also emphasizes the general importance of formats of representation when toying with theories or models. Formalisms are a specifically mathematical type of format whose role needs further investigation. This is what we do in the next section.

4. What are formalisms? As briefly stated above, formalisms are mathematical languages that allow one to present mathematical statements or objects and draw inferences about them by means of general inference rules. For example, *Hamiltonian formalism* is one of the formalisms through which scientists may find out means to solve differential equations. *Path integrals* is another formalism of this kind, with the help of which one may also solve (partial) differential equations. Let us illustrate the latter point further: the integral solution of the Schrödinger equation requires using a mathematical object, the *propagator,* whose calculation the path integrals formalism makes easier. *Fourier representation or formalism* enables one to represent mathematical functions as the continuous sum of sine functions (or complex exponential functions), so that harmonic analysis, i.e. the decomposition of a signal in its harmonic frequencies, may be performed. It also provides modelers with a way to express the solutions of some partial differential equations, such as the heat equation. Finally, formalisms like *numerical integrators, cellular automata, lattice Boltzmann methods,* and *discrete variational integrators*, are indispensable in current computational science.

Formalisms consist in the following elements:

i. elementary symbols;

ii. syntax rules that determine the set of well-formed expressions;

iii. inference rules;

iv. a partly detachable interpretation, both mathematical and physical.

Their use is facilitated by

v. translation rules that indicate how to shift from one formalism to another.

Let us illustrate these elements by discussing in more detail the above examples. In the Hamiltonian formalism, elementary symbols are used for a variable and its conju-

gate momentum: "(q, p)", or for Poisson brackets ''{.,.}''. Among the syntax rules that are specific to Hamiltonian formalism, some allow one to rewrite Hamilton equations by using the canonical variables. Inferences rules allow the users to use action-angle variables (I, *theta*) and to solve equations by using these coordinates because this change of variables opens the possibility to deal with integrable systems, thus providing a systematic method to solve *exactly*, i.e., in closed forms, differential systems like the simple pendulum, and more generally, any 1D-conservative system. Indeed, due to this change of variables, one takes full advantage of the existence of conserved quantities in mechanical systems, which are then used as variables (actions) in Hamilton equations. This allows constructing the solution of the equations by "quadrature'' (Babelon et al. 2003, chapter 2). An example of a translation rule is the Legendre transform that allows one to shift to Lagrangian formalism. Similarly, in the case of Fourier transforms, an elementary specific symbol is $f^\wedge$, which corresponds to the Fourier transform of the function *f*. Scientists use sets of rules that describe the Fourier transforms of some typical functions, such as the constant function, the unit step function, and the sinusoids, but also rules for the convolution product, viz. the Fourier transform of the convolution $f \circ g$ is the product of Fourier transforms of *f* and *g*: $(f \circ g)^\wedge = f^\wedge . g^\wedge$, so that solutions of equations may be found within Fourier space. An inverse Fourier transform is also defined, which enables one to move back from the Fourier transform $f^\wedge$ to the function *f* (this is again a translation rule).

As emphasized above, formalisms are (partly) content neutral and thus "exportable", even though they usually come with a privileged physical interpretation. As a matter of fact, most formalisms have been developed within a peculiar modeling context or are linked to a physical theory. From this origin, the most successful ones may become autonomous and depart from their original, physical interpretation. For example, Hamiltonian formalism was initially developed in the context of classical mechanics but is nowadays autonomous and used in other physical contexts. Path integrals originally come from the study of Brownian motion (Wiener 1923) and quantum mechanics (Feymann 1942) but are currently used in other fields like field theory and financial modeling.

The mathematical interpretation of formalisms may sometimes be detachable. For example, the transition rules associated with cellular automata (see below) do not have any obvious mathematical interpretation. Further, although some formalisms are linked to acknowledged mathematical theories (e.g., the Fourier formalism is linked to

the theory of complex functions), they differ from genuine mathematical theories, as shown by the example of path integrals, in which the formalism is used in the absence of any uncontroversial mathematical theory that could back it up. The definition of a path integral:

$$K(b,a) = \int_a^b e^{\frac{2i\pi}{h}\int_{t_b}^{t_a} L\,dt}\, Dx(t)$$

requires using a measure "Dx", to which no general, rigorous definition can be given yet. This mathematical concern does not prevent physicists from using path integrals anyway, as testified by the following quote: "The question of how the path integral is to be understood in full generality remains open. Given this, one might expect to see the physicists expending great energy trying to clarify the precise mathematical meaning of the path integral. Curiously, we again find that this is not the case" (Davey 2003, 450).

Let us finally emphasize that formalisms also differ from formulations of physical theories and allow philosophers of science to address different philosophical problems. Formulations of theories, in particular axiomatic ones, are explored when questions about conceptual content and metaphysical implications are raised. They pertain to foundational issues. Whether a given formulation involves calculus is a peripheral issue in this context. By contrast, the primary virtue of a formalism is to allow modelers to draw actual inferences from a theory or model. The inferential rules it contains are more important than the mathematical rigor of the language in which it is expressed.

5. Choosing a formalism. So far, we have argued that the inferential power that is required to explore models is partly brought about by formalisms, and we have given examples thereof. Accordingly, formalisms have to be carefully examined by philosophers of science if they are to provide a fine-grained analysis of how scientific knowledge is produced in practice. We now aim to show that there is no unique description of formalism-rooted inferential power since different formalisms allow for different types of inferences and are adapted to different types of inquiries. We do so by providing examples of these differences and of the factors that guide scientists when choosing the formalism that is best suited to the task at hand.

How do scientists decide which formalism to use in a given inquiry? The choice may first depend on the type of models at hand. For example, the path integral formalism is

well adapted to solve systems with many degrees of freedom (Zinn-Justin 2009) and makes ''certain numerical calculations in quantum mechanics more tractable'' (Davey 2003, 449). Lagrangian formalism offers a well-suited framework to solve equations describing constrained systems (Goldstein 2002, 13, Vorms 2009, 15). Fourier representation allows one to solve, e.g., the differential equations describing the time evolution of electrical quantities in networks. In this case, differential equations are transformed into *algebraic equations* on variables in Fourier space, which may be easier to solve. Finally, with the change of action-angle variables, Hamiltonian formalism potentially provides exact solutions for integrable systems, which have as many independent conserved quantities as degrees of freedom.

The use of a particular formalism is also guided by epistemic goals. Depending on the chosen formalism, different kinds of properties, general (e.g. periodicity, symmetry) or particular (dynamical), may be inferred from the same model. Let us illustrate this point with the example of prey-predator models in ecology. Among these, some obey Lotka-Volterra (LV) equations and represent transforming populations with a system of two coupled equations. If they are investigated within the Hamilton formalism, *general properties* of these models can be found without setting initial conditions or numerical values for the involved parameters. The reframed models can indeed be shown to be integrable, like the simple pendulum in classical mechanics. Dutt explicitly emphasizes the advantages of using this formalism for a two-species LV system:

> ''In dealing with the problems involving *periodicity*, the Hamilton-Jacobi canonical theory has a distinct advantage over the conventional methods of classical mechanics. In this approach, one introduces action and angle variables through canonical transformations in such a way that the angle variable becomes cyclic. One then obtains the frequency of oscillation by taking the derivative of the Hamiltonian with respect to the action variable. One may thus *bypass the difficulty* in obtaining the complete solutions of the equations of motion, *if these are not required*.'' (Dutt, 1976, 460, our emphasis)

LV models can also be solved with the help of computers and generic numerical integrators when the aim is to obtain particular dynamics for specific values of parameters and initial conditions. Such numerical solutions of the LV model can also be provided by specific formalisms, such as discrete variational integrators (Krauss 2017, 34; Tyranowski 2014, 149). In that case, discrete equations are derived from a discrete least action principle, which is well-suited to conservative systems, like the LV sys-

tem. Discrete variational integrators allow for the preservation of general properties like the conservation of global quantities, viz. energy, momenta, and symplecticity. This discrete formalism comes with mathematical constraints on the discretization of time since the time step has to be adaptive in order to guarantee the conservation of global quantities (Marsden & West 2001, Section 4.1).

Finally, let us mention that LV models can also be studied by using *cellular automata* (CA) and associated formalism, with the following advantages:

> [a rather general predator-prey model] is formulated in terms of automata networks, which describe more correctly the *local character* of predation than differential equations. An automata network is a graph with a discrete variable at each vertex which evolves in discrete time steps according to a definite rule involving the values of neighboring vertex variables. (Ermentrout and Edemstein-Keshet 1993, 106)

On the one hand, CA are discrete dynamical systems, but on the other, they are also a nice means to practice science with the help of a computationally simple formalism (in terms of transition rules). They can be extremely powerful. For example, rule 110 is Turing complete and, like lambda-calculus, can emulate any Turing machine and therefore complete any computation. In contrast with the case of Hamilton formalism, CA-based inferences from prey-predator models are carried out for specific values and parameters. As CA are described by local rules, these inferences merely pertain to local variations in the model. However, the simplicity of these rules is a tremendous advantage for modeling and code-writing. For instance, CA allow one to easily add rules for the pursuit and evasion of populations as well as rules for age variation (Boccara et al. 1993, Ermentrout and Edemstein-Keshet 1993, see also Barberousse and Imbert 2013 for an analysis of CA as used in fluid dynamics and compared with Navier-Stokes based methods).

Let us now turn to a different example illustrating how different the epistemological effects of using this or that formalism may be. Crystals are currently modeled as lattices that come under two forms, *lattices in real space* and *lattices in reciprocal space*. Each is associated with a specific formalism. Within the *real space lattice* formalism, crystals are described with a vector R expanded on a vector basis (a_1, a_2, a_3) which corresponds to crystal directions, and *alpha*, *beta*, *gamma* are the corresponding angles. Inferences about *symmetry* of crystals are usually made within this type of representation since the real space is well adapted to studying discrete translations and rotations.

Crystals can also be described with the help of a vector R* in a *lattice in reciprocal space*. There is a clear correspondence between the two spaces since they are dual. Given R in the real space, we can derive R* in the reciprocal space, and conversely. The two spaces are related by a Fourier transform. However, the *reciprocal space* can be more convenient because inferences about *diffraction and interference patterns* are easier to carry out in the Fourier representation. As stressed by Hammond in a textbook of crystallography:

> the reciprocal lattice is the basis upon which the geometry of X-ray and electron diffraction patterns can be most easily *understood* and […] the electron diffraction patterns observed in the electron microscope, or the X-ray diffraction patterns recorded with a precession camera, are simply sections through the reciprocal lattice of a crystal (Hammond 2009, 165).

This example shows that facilitating inferences may have various epistemological effects. Some are relevant to computational aspects and the predictions or explanations that scientists are able to produce in practice. Others pertain to the way scientists understand and reason about models and their target systems. This example also shows how different epistemic goals (symmetry-oriented vs. interference-oriented investigations of crystals) determine which formalism is chosen.

Overall, the above shows that formalisms not only have an important impact on the amount of results scientists may produce, but also on the types of results that are attainable. The examples we have discussed also highlight that the existence of a variety of formalisms is a source of epistemic richness and enhanced inferential power for scientists because it provides them with multiple ways of investigating the same mathematical structures or structures that are related by suitable morphisms.

6. Conclusion. The above proposals are meant to contribute to the epistemological question of what provides models with inferential power and helps scientists succeeding in their inquiries. We have shown that some of this inferential power is brought about by the formal symbolic tools that scientists use to present and investigate mathematical models. Our second claim is that all formal settings do not enable the same types of inferences nor are suited to all epistemic goals. Accordingly, a fine-grained analysis of the conditions of scientific progress needs, among other things, to focus on formalisms.

Our epistemological analysis is not tied to any particular theory of scientific representation. However, by showing that inferences actually hinge on choice of formalism, it suggests that a theory of scientific representation that is cashed out in terms of structures is too abstract to account for the various ways equations are solved in practice and information extracted from scientific models.

References

Babelon Olivier, Bernard Denis, and Talon Michel. 2003. *Introduction to classical integrable systems*, Cambridge: Cambridge University Press.

Baker, A. 2009. "Mathematical Explanation in Science". *British Journal for the Philosophy of Science* 60 (3): 611–633.

Barberousse, Anouk, and Cyrille Imbert. "New Mathematics for Old Physics: The Case of Lattice Fluids." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 44 (3) : 231–41.

Barberousse, Anouk, and Cyrille Imbert. 2014. "Recurring Models and Sensitivity to Computational Constraints" *The Monist* 97 (3): 259–79.

Boccara Nino, Roblin O. and Roger Morgan. 1994. Automata network predator-prey model with pursuit and evasion, *Physical Review E* 50 (6): 4531–41

Bueno, Otávio, and Mark Colyvan. 2011. "An Inferential Conception of the Application of Mathematics". *Noûs* 45 (2): 345–74.

Bueno, Otávio. 2014. "Computer Simulations: An Inferential Conception". *The Monist* 97 (3): 378–98.

Cartwright, Nancy (1983). *How the Laws of Physics Lie*. Clarendon Press, Oxford.

Cartwright, Nancy.1999. "Models and the Limits of Theory: Quantum Hamiltonians and the BCS Models of Superconductivity". In *Models as Mediators*, ed. Mary S. Morgan and Margaret Morrison Morgan, Cambridge: CU Press: 241–81.

Colyvan, Mark. Forthcoming. "The Ins and Outs of Mathematical Explanation", *Mathematical Intelligencer*.

Davey Kevin. 2003. "Is Mathematical Rigor Necessary in Physics?" *The British Society for the Philosophy of Science*, 54(3): 439–463

Dutt Ranabir. 1976. "Application of the Hamiltonian-Jacobi Theory to Lotka-Volterra Oscillator", *Bulletin of Mathematical Biology*, 38: 459–465.

Ermentrout G. Bard and Edemstein-Keshet, Leah. 1993. "Cellular Automata Approaches to Biological Modeling". *Journal of Theoretical Biology* 160: 97–133.

Feynman, Richard. P. 1942. "The Principle of least action in quantum mechanics", *PhD. diss.*, Princeton University.

Frigg, Roman. 2010. "Models and Fiction". *Synthese* 172 (2): 251–68.

Frigg, Roman, and Stephan Hartmann. 2017. "Models in Science." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/models-science/.

Goldstein, Herbert. 2002. *Classical Mechanics*. Reading, Mass: Addison-Wesley.

Hammond, Christopher. 2009. *The Basics of Crystallography and Diffraction*, Oxford University Press.

Hughes, Robert I.G. 1997. "Models and Representation". *Philosophy of Science* (Proceedings): 64: S325–S336.

Hughes, Robert I.G. 2010. *The Theoretical Practices of Physics: Philosophical Essays*. Oxford: Oxford University Press.

Humphreys, Paul. 2004. *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Kraus, Michael. 2017. "Projected Variational Integrators for Degenerate Lagrangian Systems", preprint: https://arxiv.org/pdf/1708.07356.pdf

Marsden Jerrold E. and West Matthew. 2001. "Discrete Mechanics and Variational Integrators", *Acta Numerica*, 10: 357–514.

Morgan, M., and Margaret Morrison (1999). *Models as Mediators*. Cambridge University Press.

Pincock, Christopher. 2004. "A new perspective on the problem of applying mathematics", *Philosophia Mathematica* 3 (12), 135-61.

Redhead, M. 1980. "Models in Physics", *The British Journal for the Philosophy of Science*, 31(2): 145-163

Suarez, Mauricio. 2002. "An Inferential Conception of Scientific Representation", *Philosophy of Science* 71 (5): 767-779

Tyranowski Tomasz. M. 2014. "Geometric integration applied to moving mesh methods and degenerate Lagrangians". Ph.D. diss., California Institute of Technology.

Vorms, Marion. 2011. "Formats of Representation in Scientific Theorizing." In *Models, Simulations, and Representations*, edited Paul Humphreys and Cyrille Imbert. Routledge.

Wiener, Norbert. 1923. "Differential space". *Journal of Mathematical Physics* 2: 131-174.

Zinn-Justin Jean. (2009), Path Integral, *Scholarpedia*, 4(2): 8674.

Representation Re-construed: Answering the Job Description Challenge with a Construal-based Notion of Natural Representation

Abstract: Many philosophers worry that cognitive scientists apply the concept REPRESENTATION too liberally. For example, William Ramsey argues that scientists often ascribe natural representations according to the "receptor notion," a causal account with absurd consequences. I rehabilitate the receptor notion by augmenting it with a background condition: that natural representations are ascribed only to systems construed as organisms. This Organism-Receptor account rationalizes our existing conceptual practice, including the fact that scientists in fact reject Ramsey's absurd consequences. The Organism-Receptor account raises some worrying questions, but as a more faithful characterization of scientific practice it is a better guide to conceptual reform.

Abstract: 100 words

Total: 4,995 words

1. Introduction. There is a common complaint among philosophers that scientists use the word "representation" too liberally. Representation is often contrasted with indication: representation is a distinction achieved by maps, linguistic performances, and thoughts, whereas indication is a less-demanding state achieved by thermostats, which indicate ambient temperature, and refrigerator lights, which indicate whether the door is open (Dretske 1981; Cummins and Poirier 2004). However, cognitive scientists often ascribe representations when it seems that mere indication is all that is called for. We commonly say that hidden layers in a neural network represent concepts, or that neurons in V1 represent visual edges, because they reliably respond differently to the circumstances they are said to represent (Ramsey 2007, 119–20; cf. Hubel and Wiesel 1962). But these "representations" are thin-blooded compared to paradigmatic conventional representations. For example, they cannot be invoked in the absence of an appropriate stimulus. So are cognitive scientists conceptually confused? Do they exaggerate their claims? And if the natural representations posited by cognitive scientists aren't genuine representations, is the cognitive revolution dead?

William Ramsey provides an excellent book-length exploration of these worries, articulating a qualified pessimism about their answers:

> …we have accounts that are characterized as "representational," but where the structures and states called representations are actually doing something else. This has led to some important misconceptions about the status of representationalism, the nature of cognitive science and the direction in which it is headed. (2007, 3)

Ramsey describes the "job description challenge": to give an account of the distinctive properties of representations in virtue of which appealing to them serves a special

explanatory role. If the job description challenge can be met, then we can formulate a plan for conceptual reform.

I undertake Ramsey's challenge, but with a metadiscursive twist: I describe the Organism-Receptor account, which articulates conditions for ascribing representations, in virtue of which such ascriptions achieve a special explanatory purpose. The account is merely suggestive about the properties that distinguish first-order representational states from non-representational states; it says more about the mental state of the ascriber than about the representation-bearing system. However, the Organism-Receptor account provides a more adequate characterization of scientists' practice than Ramsey's.

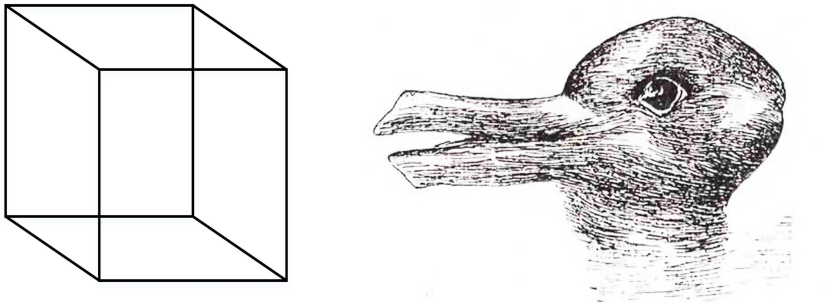
My main aim in this paper is to push back against pessimistic evaluations of the existing practice of representation-ascription in cognitive science, like Ramsey's. I will focus on Ramsey's critique of the "receptor notion," a flawed causal theory of representation that he attributes to some cognitive scientists. Ramsey argues that the receptor notion has absurd consequences, although scientists do not accept them. By augmenting the receptor notion with a construal-based background condition, I can explain why scientists do not draw these absurd conclusions. Whereas Ramsey's pessimistic account of scientists' practice of ascribing representations finds it wanting and is extensionally inadequate, mine rationalizes our extant conceptual practice (though that practice is not beyond criticism). I conclude that my apologetic account is a more charitable and adequate interpretation of existing scientific practice than Ramsey's.

2. Ramsey on the "Receptor Notion." Ramsey argues that natural representations in cognitive science are often ascribed according to the "receptor notion," a crude causal theory of representation. According to the receptor notion, a state *s* represents a state of affairs *p* if *s* is regularly and reliably caused by *p* (2007, 119).

Ramsey claims that the receptor notion is what justifies the ascription of representations to cells in V1 that detect visual edges, cells in frog cortex that detect flies, and the mechanisms in Venus flytraps that cause their "jaws" to close (119–23). Ramsey argues that this receptor notion is too liberal to be useful to scientists. For example, it is susceptible to the "disjunction problem" (Fodor 1987): since frog neurons respond reliably to visual stimulation by flies *or* (say) BBs, we should say that the content of the representation is *fly-or-BB*, rather than *fly*. Likewise, Venus flytraps represent objects in a particular range of sizes rather than *edible insects*, and the human concept GOAT represents *goats-or-weird-looking-sheep*. Such disjunctive content-ascriptions are usually considered absurd. Absent a clever fix, we must embrace unwieldy, disjunctive contents for representations or we must reject the receptor notion (Ramsey, 129).

Dretske's (1988) teleofunctional theory of representation is a sophisticated twist on the receptor notion that avoids the disjunction problem. On Dretske's view, a representational state must not only be causally dependent on the state of affairs it represents, but must serve a function for its containing system in virtue of this causal dependency. This extra condition motivates constraints on representational content that eliminate problematic disjunctive contents. Dretske's theory is subject to some subtle criticisms that I will discuss in Section 6, but the Organism-Receptor account will preserve some of the teleological character of Dretske's theory.

Ramsey's most compelling objection to the receptor account, including Dretske's sophisticated version, is that it justifies ascribing representational contents to states that are not, in fact, representational: smoke "represents" fire since the latter causes the former. Likewise, the firing pin of a gun "represents" whether the trigger is depressed, and rusting iron "represents" the presence of water and oxygen (138–47). Ramsey claims, plausibly, that these are absurd consequences. I find Ramsey's reductio



Ambiguous figures. Left: The Necker cube. Right: The duck-rabbit (image from Jastrow 1899).

compelling, but reject a different premise than he does. Rather than conclude that cognitive scientists have a bad conceptual practice, I question whether his characterization of the receptor notion is a charitable understanding of what happens in cognitive science. After all, cognitive scientists do not generally claim that GOAT denotes *goats-or-sheep* (at least for competent judges of goathood), or that firing pins represent anything.

3. A Construal-based Notion of an Organism. I argue that something like the receptor notion can be salvaged if being a receptor is contextualized in terms of construal. Construal (also called "seeing-as") is a judgment-like attitude whose semantic value can vary licitly independently of the state of affairs it describes. For example, we can construe an ambiguous figure like the Necker cube as if it were viewed from above or below, or the duck-rabbit as if it were an image of a duck or of a rabbit (Roberts 1988; see also Wittgenstein 1953). We can construe an action like

skydiving as brave or foolhardy, depending on which features of skydiving we attend to.

On a construal-based account of conceptual norms, a concept (e.g. REPRESENTATION) is ascribed relative to a construal of a situation. For example, perhaps I fear something only if I construe it as dangerous to me or detrimental to my ends (Roberts 1988). Daniel Dennett's (1987) intentional stance is a more familiar example: according Dennett, a system has mental states if and only if we construe it in such a way that its behavior is explainable in terms of a belief-desire schema.

I propose that construing something as an organism involves construing it such that it has goals and behavior, and believing that it has mechanisms that promote those goals by producing that behavior. More precisely:

> **Organism-Construal.** A subject *a* construes a system *x* as an organism in a context[1] *c* if and only if, in *c*,
>
> (O1) *a* attributes a set of goals *G* to *x*,
>
> (O2) *a* attributes a set of behaviors *B* to x,
>
> (O3) *a* believes that the elements of *B* function to promote elements of *G*,
>
> (O4) *a* believes that *x* possesses a set of mechanisms *M*, and
>
> (O5) *a* believes that the elements of *M* collectively produce the elements of *B*.

My main argument does not rely on all the details of Organism-Construal; it could be replaced by a different explication of what it is to see something as an organism. But Organism-Construal captures an intuitive notion of a critter. First of all, we normally take living critters to have goals, such as survival and reproduction, and behaviors that

[1] The relevant notion of a context is something like MacFarlane's (2014) "context of assessment."

promote those goals. However, Organism-Construal does not require that an organism really have goals (whatever that involves) or exhibit behavior (however that's distinguished from other performances). To see something as an organism according to Organism-Construal, the construing subject need only *attribute* goals to the system, and see some of its performances as behaviors that promote those goals. Such goals could include relatively specific aims such as locating food, getting out of the rain, or driving home. We sometimes also attribute goals and behaviors to non-living things, such as automated machines. For example, we might say that a robot vacuum has the goal of cleaning the floor, which it accomplishes by sucking up dust. Or I might say that my GPS navigation computer is trying to kill me, which it accomplishes by consistently giving me directions that lead me through strange, dangerous backroads. Condition (O3) is expressed in terms of belief instead of attribution, meaning that the construing subject must sincerely believe that an organism's putative behaviors function to promote its putative goals. When and insofar as someone construes a system in this way, the conditions (O1)–(O3) above are satisfied.

Conditions (O4)–(O5) require that the system's behavior be explainable by appeal to mechanisms. "Mechanisms" here should be understood in roughly the sense meant by the new mechanists (Machamer, Darden, and Craver 2000; Bechtel and Abrahamsen 2005; Craver 2007): organized structures of component parts and operations that produce a phenomenon, and the description of which is an explanatory aim of some scientific projects. Much explanation in biology and neuroscience plausibly follows a mechanistic model, and likewise in cognitive science. Daniel Weiskopf (2011) has argued that cognitive explanations are not properly mechanistic, but even on his view cognitive explanations are extremely similar to mechanistic ones, distinguishable only because the relationship between components of cognitive models and their physiological realizers is relatively opaque. Regardless, cognitive scientists use the word "mechanism" to refer to the referents of their models,

just as biologists and neuroscientists do. I am more moved by the similarities between the biological and the cognitive sciences than the differences. Therefore, like Catherine Stinson (2016), I acknowledge Weiskopf's concerns but nevertheless adopt the language of "mechanisms."

Not all of a system's mechanisms function to produce behavior. For example, biological organisms have metabolic and other mechanisms that maintain bodily integrity. Such mechanisms may need to function correctly as a background condition for the organism to behave, but scientists do not typically take behavioral patterns to be the explanandum phenomena of such mechanisms. Let us call mechanisms that do contribute to the explanation of behavior *behavioral mechanisms*. As for what it means for a system to "possess" a mechanism, a mereological criterion will do for now: the mechanism must be a part of the system. Condition (O5) is meant to limit the mechanisms in the set *M* to behavioral mechanisms.

So far so abstract; let's consider an example. The robot Herbert was designed to wander autonomously through the MIT robotics lab, avoiding obstacles, and collecting soda cans with its arm (Brooks, Connell, and Ning 1988). Herbert can be construed as an organism, even though it is not alive, as long as one (O1) attributes goals, like avoiding collisions and collecting soda cans, to Herbert, (O2) sees some of Herbert's performances as behaviors, (O3) believes that Herbert's behaviors promote its goals, and (O4) believes that Herbert possesses mechanisms that (O5) explain its behavior. Herbert does possess mechanisms for accomplishing goals; it is equipped with sensors, computers, and motors that coordinate its locomotion and its grasping arm. And most people readily anthropomorphize Herbert enough to see it as a goal-directed, behaving system (pace Adams and Garrison [2013], who insist that Herbert has its designers' goals, but no goals of its own). Anyone willing to engage in the imaginative attribution of goals and behavior to Herbert can see Herbert as an organism, even if on reflection they believe Herbert is not literally an organism. The

willingness to ascribe representations to a system plausibly waxes and wanes along with one's willingness to construe the system as an organism in something like the sense described above. There are psychological limits on the willingness to attribute goals and behaviors to systems relatively unlike animals, and these limits may vary between individuals.

4. The Receptor Notion Re-construed. Returning now to the receptor notion of natural representation, I suggest that it can be augmented in the following way:

> **Organism-Receptor.** A state *s* represents a state of affairs *p* if
>
> (R1) *s* is regularly and reliably caused by *p*, and
>
> (R2) *s* is a functional state of a behavioral mechanism possessed by an organism.

Organism-Receptor is not a construal-based explication, but it depends on a construal-based account of ORGANISM. It preserves the spirit of Ramsey's receptor notion, with the added condition that representations be ascribed to parts of systems construed as organisms. Representation-ascriptions guided by Organism-Receptor inherit their plausibility from the plausibility of the corresponding construal of some system as an organism. Most accounts of cognitive representation require there to be a representational subject of some kind (e.g. Adams and Aizawa 2001; Rupert 2009; Rowlands 2010), and on Organism-Receptor the organism serves this role. We can constrain the acceptable contents of these representations by requiring they correspond to descriptions of *p* according to which *p* is relevant to the pursuit of an organism's goals. This appeal to goals is not ad hoc, since according to Organism-Receptor representations are ascribed to organisms, i.e. systems to which we've already attributed a set of goals. Thus, like Dretske's (1988) and Millikan's (1984)

teleofunctional accounts, this construal-based account addresses the disjunction problem by appealing to goals of organisms.

The metadiscursive job-description challenge is to provide criteria of ascription for representations, in virtue of which representation-ascriptions achieve some explanatory purpose. I have provided criteria of ascription, so what is their purpose? On Donald Davidson's (1963, 5) account of intentional action, actions are performed under the guise of a privileged description (or set of descriptions). Davidson flips the light switch in order to turn on the light, but not in order to alert the prowler outside (whose presence is unknown to Davidson) that he is home, though he also does the latter. Davidson calls this feature of action its "quasi-intensional character." Behavioral mechanisms also have something like a quasi-intensional character, since there are privileged descriptions that make explicit how they and their components contribute to an organism's capacity to pursue its goals. For example, edge-detecting cells in V1 fire in order to identify boundaries in an organism's environment, not to consume glucose, though they also do the latter. The use of representation-talk by cognitive scientists, as licensed by Organism-Receptor, is a way to habitually mark these privileged descriptions and distinguish them from other descriptions of the same states or events. And since cognitive science is concerned with the functional structure of behavior-coordinating mechanisms rather than other features of cognitive systems, it is easy to see why representation—even in this relatively thin sense—has always been the dominant theoretical perspective in cognitive science. This focus on quasi-intensional characterization may even be what makes the cognitive scientific perspective distinctive (on scientific perspectives, see e.g. Giere 2006).

The Organism-Receptor account provides us with resources to salvage the receptor notion from Ramsey's reductio. It is plausible to suppose that cognitive scientists generally ascribe natural representations to systems against an imaginative

background like this. After all, most cognitive science concerns the mechanisms of living systems, especially animals (except in computer science and some computational modeling, where the object of attention is a formal object like a connectionist network that is presumed to be analogous in some way to such a mechanism). Such systems are easily construed as organisms in the sense of Organism-Construal. Non-living things and even non-animals are in general more difficult to construe as organisms in that sense, since they are often perceived to lack goals, the capacity to behave, or both.

5. The Organism-Receptor Notion in Context. Consider a strong case of representation, like fly-detecting cells in frog visual cortex. We construe frogs as systems that exhibit goal-directed behavior and believe they possess mechanisms that explain that behavior. Frog visual cortex contains mechanisms that (along with other mechanisms) explain behaviors like fly-catching. When we identify cells in frog visual cortex that fire in response to the visual presence of flies (or fly-like objects), we ascribe representational properties to those cells. The contents we ascribe to representations in frog visual cortex are constrained by the goals we attribute to frogs. *That a small insect is present* is a suitable content because flies can be consumed for energy; *that a wiggly BB is present* does not have this significance for frogs, although BBs may be indistinguishable from insects by the mechanisms in the frog's visual cortex. Nevertheless, the relationship between fly-presence and the frog's goals provide a ground for privileging non-disjunctive descriptions of representational content.

The Organism-Receptor account also explains why liminal cases of representation, like the case of Herbert, are liminal. We can say that Herbert represents such states of affairs as the presence of obstacles and soda cans, because states of Herbert's sensors are regularly and reliably caused by those states of affairs.

And we can ascribe contents to representations by drawing on descriptions of Herbert's environment that relate to the goals we ascribe to Herbert. However, our willingness to take these representations seriously as natural representations that bear content intrinsically covaries with our willingness to take Herbert seriously as an organism. We are not as comfortable attributing genuine goals and behaviors to Herbert as we are attributing goals and behaviors to frogs.[2]

Finally, absurd cases like the firing pin can be excluded (for the most part) since guns are not easily construed as "organisms." Firearms are difficult to anthropomorphize, since they do not exhibit autonomous behavioral dynamics and we don't normally see them as having goals of their own. It is not *impossible* to ascribe goals to weapons or other tools, but the ascription of folk-psychological properties to tools, like the folk ascription of a bloodthirsty disposition to a sword, generally depends on the way a tool influences its users' behavior. (I suspect this dependence might offer some novel explanations of why Clark and Chalmers' [1998] extended cognition hypothesis is attractive to some.) The attribution of autonomous behaviors to tools like swords is fanciful. Perhaps we might imagine a tool exhibits psychic "behavior," but anyway we do not believe that swords possess mechanisms that produce this "behavior" (though if we did, such a construal would be more compelling). If the firing pin of a gun is not a component of a behavioral mechanism, it cannot represent anything according to the Organism-Receptor account.

So the Organism-Receptor account licenses an ascriptive practice that resembles the crude receptor notion when the role of construals is not made explicit. It is unusual in that it inverts Ramsey's preferred order of ascription: Ramsey wishes to

---

[2] Notably, Rodney Brooks himself does not claim that it is proper to ascribe representational capacities to Herbert (Brooks, Connell, and Ning 1988; Brooks 1991), but Brooks plausibly had in mind a more demanding account of representation.

ascribe cognitive structure to systems in virtue of their representational structure (see e.g. Ramsey, 222–235), whereas I suggest that we in fact ascribe representational structure in virtue of seeing a system as a system with goal-directed behavior, i.e. as a potentially cognitive system.

6. Worries. Since the Organism-Receptor account shares a certain teleological character with Dretske's account, I will discuss Ramsey's two most developed objections to Dretske, along with other worries specific to the Organism-Receptor account. First, Ramsey objects that Dretske's account is question-begging with regard to the job-description challenge. Roughly, teleological normativity (i.e. functioning and malfunctioning) is not sufficient to explain intentional normativity (i.e. representation and misrepresentation), and since Dretske provides no satisfying criteria for what it is for a state to function as a representation, he cannot bridge that gap (Ramsey 2007, 131–2). But the Organism-Receptor account has more resources than Dretske's teleofunctionalism. Construing a system as an organism involves construing it as exhibiting behavior, which allows us to distinguish behavioral mechanisms from other mechanisms. On the Organism-Receptor account, misrepresentations are malfunctions of behavioral mechanisms (like frog vision), but not of other mechanisms (like a frog's circulatory system or a gun's firing mechanism).

My reply invites a rejoinder: on the Organism-Receptor account the functional roles of representations will be extremely diverse, and representations will be common. They will not just include IO-representation and S-representation (roughly, information-processing relata and models for surrogative reasoning; Ramsey 2007, 68ff.), which Ramsey and most cognitive scientists regard as genuinely representational. They will also include more controversial varieties of "representation," such as Millikan's (1995) "pushmi-pullyu" representations: Janus-faced mechanistic components that simultaneously indicate a state of affairs and cause

an adaptive or designed response. In other words, representations will include what Ramsey calls "causal relays" like the firing pin in a gun, the inclusion of which in the extension of REPRESENTATION was the ground for his reductio! However, the absurd cases can be avoided. The firing pin case is excluded because guns are poor examples of organisms. And pushmi-pullyu representations include cases with significant intuitive appeal to many scientists, like the predator calls of vervet monkeys (Millikan 1995; cf. Seyfarth, Cheney, and Marler 1980). While this conception of representation has a more liberal extension than Ramsey is comfortable with, it is liberal enough to explain common representation-ascriptions in cognitive science without being so liberal as to countenance absurd cases like Ramsey's firing pin, so I submit it is adequate to scientific practice.

Ramsey's second objection is that Dretske is committed to a false principle: that if a component is incorporated into a mechanism because it carries information, then its function is to carry information (132–9). However, the Organism-Receptor account constrains the causal dependence criterion (R1) by relying on construals of systems as organisms instead of teleofunctional commitments. The account I describe is not committed to Dretske's principle, and therefore is not subject to this objection.[3]

Nevertheless, one might worry whether the organism criterion (R2) is a suitable condition on representation-ascription. I suggested five conditions (O1)–(O5) on what can be seen as an organism, but conditions (O1) and (O2) are fairly unconstrained. There are psychological limitations on when goals or behaviors can be plausibly attributed to a system, but what are those limits? And what factors influence interpersonal variability in willingness to make these attributions? The reason this practice isn't bonkers is that it coheres with the explanatory purpose of

[3] Ramsey's discussion is rich and worthy of deeper engagement than this, but for reasons of space I leave the matter here.

representation-ascriptions: to make explicit the quasi-intentional character of behavioral mechanisms. Nevertheless, we should hope that these psychological limitations are vindicated by more principled considerations. Criticism is warranted if scientists attribute goals and behaviors when they should not. There is some extant work on the proper norms ascribing goals to organisms (e.g. Shea 2013; Piccinini 2015, chap. 6), but little serious work on how to understand the concept of BEHAVIOR in the context of cognitive science. We should worry about the practice of ascribing natural representations if scientists construe things that are not cognitive systems as "organisms." Indeed, we might indeed worry that many cognitive scientists misuse the concept COGNITION, given the intense disagreements over its extension (see e.g. Akagi 2017). However, my present aim is not to evaluate scientific practice, but to describe it faithfully (with the hope that a more satisfactory evaluation will follow).

Another worry about construal-based accounts is that they entail an unattractive anti-realism: if representations and their contents only exist relative to construals, they are mind-dependent rather than objective, right? This worry is unfounded. I am undertaking a modified version of Ramsey's job description challenge: my aim is to describe the ascription of representations in virtue of which they serve an explanatory purpose, not to distinguish genuinely representational states from non-representational states. The Organism-Receptor account does not entail that representations exist relative to construals, only that they are *ascribed* relative to construals. My account is consistent with the existence of a first-order account of the metaphysics of representation that justifies this practice (or doesn't). After all, the duck-rabbit can be construed as a duck even if it is not a duck, and nothing about that fact entails that ducks (or unambiguous images of ducks) are not real. The Organism-Receptor account describes a norm that plausibly guides human scientists with imperfect capacities for knowledge. But while my solution to the metadiscursive job description challenge is not inconsistent with Ramsey's solution to

the first-order job description challenge, it is inconsistent with Ramsey's characterization of scientific norms for ascribing natural representations.

7. Conclusion. I began by observing the common worry that scientists ascribe representations more liberally than many philosophers are comfortable with, and in particular that scientists rely on an unsatisfactory "receptor" criterion. I sketched an account on which scientists ascribe natural representations only to components of mechanisms of systems construed as "organisms." Since in practice cognitive scientists attend almost exclusively to systems that are easily so construed, their behavior may appear to be guided by the crude receptor criterion whereas in fact it is guided by the Organism-Receptor criterion. However, while the Organism-Receptor account is still relatively liberal, a crucial difference between the two accounts is that the crude criterion has absurd consequences, whereas such consequences are eliminated or marginalized on the Organism-Receptor criterion. Since scientists do not in fact endorse these absurd consequences, I argue that the augmented criterion is a better hypothesis regarding norms for representation-ascription in cognitive science.

This is proposal is not a comprehensive, new theory of representation, but it accomplishes two things. First, it provides argumentative resources for resisting the common worry that cognitive scientists use hopelessly liberal criteria for ascribing representations. Second, it offers a novel picture of practices for representation-ascription in the biological and behavioral sciences, one that is less pessimistic picture than Ramsey regarding conceptual rigor in cognitive science. The picture is not beyond criticism—in particular, it wants for a more detailed account of the grounds that warrant attributing behaviors and goals to systems. But since it is more faithful to our practice than Ramsey's it is likely to yield more productive suggestions for how to guide that practice into the future. I suggest that we safeguard conceptual rigor in cognitive science not by cleaving more faithfully to the representationalism of the

cognitive revolution, but by embracing role of construal in scientific inquiry, making it explicit, and subjecting it to reasoned criticism.

REFERENCES

Adams, Fred, and Ken Aizawa. 2001. "The Bounds of Cognition." *Philosophical Psychology* 14:43–64.

Adams, Fred, and Rebecca Garrison. 2013. "The Mark of the Cognitive." *Minds and Machines* 23:339–52.

Akagi, Mikio. 2017. "Rethinking the Problem of Cognition." *Synthese*. doi: 10.1007/s11229-017-1383-2.

Bechtel, William, and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421–41.

Brooks, Rodney. 1991. "Intelligence without Representation." *Artificial Intelligence* 47:139–59.

Brooks, Rodney, Jonathan Connell, and Peter Ning. 1988. "Herbert: A Second Generation Mobile Robot." *A.I. Memos* 1016:0–10.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58:7–19.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Cummins, Robert, and Pierre Poirier. 2004. "Representation and Indication." In *Representation in Mind: New Approaches to Mental Representation*, E sds. Hugh Clapin, Phillip Staines and Peter Slezak, 21–40. Amsterdam: Elsevier.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* 60:685–700.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

———. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT/Bradford.

Giere, Ronald N. 2006. *Scientific Perspectivism*. Chicago: University of Chicago Press.

Hubel, David H., and Torsten N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160:106–54.

Jastrow, Joseph. 1899. "The Mind's Eye." *Popular Science Monthly* 54:299–312.

MacFarlane, John. 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Clarendon.

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67:1–25.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

———. 1995. "Pushmi-Pullyu Representations." *Philosophical Perspectives* 9:185–200.

Piccinini, Gualtiero. 2015. *Physical Computation: A Mechanist Account*. Oxford: Oxford University Press.

Ramsey, William M. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.

Roberts, Robert C. 1988. "What Emotion Is: A Sketch." *Philosophical Review* 97:183–209.

Rowlands, Mark. 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: MIT Press.

Rupert, Robert. 2009. *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.

Seyfarth, Robert M., Dorothy L. Cheney, and Peter Marler. 1980. "Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication." *Science* 210:801–3.

Shea, Nicholas. 2013. "Naturalising Representational Content." *Philosophy Compass* 8:496–509.

Stinson, Catherine. 2016. "Mechanisms in Psychology: Ripping Nature at Its Seams." *Synthese* 193:1585–614.

Weiskopf, Daniel A. 2011. "Models and Mechanisms in Psychological Explanation." *Synthese* 183:313–38.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. 3rd Ed. Trans. G.E.M. Anscombe. Eds. G.E.M. Anscombe and Rush Rhees. Oxford: Blackwell, 2001.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -122-

Comparing Systems Without Single Language Privileging

Max Bialek

mbialek@rutgers.edu

For the 2018 PSA Meeting.

Word count: 4753

Abstract

It is a standard feature of the BSA and its variants that systematizations of the world competing to be the best must be expressed in the same language. This paper argues that such single language privileging is problematic because (1) it enhances the objection that the BSA is insufficiently objective, and (2) it breaks the parallel between the BSA and scientific practice by not letting laws and basic kinds be identified/discovered together. A solution to these problems and the ones that prompt single language privileging is proposed in the form of privileging the best system competition(s).

1 Introduction

According to the Best Systems Analysis (BSA), the laws of nature are the theorems of the best systematization of the world—with 'best' standardly understood to mean the simplest and most informative (on balance). It is currently a standard feature of the BSA (since Lewis 1983) and its variants (Loewer 2007; Schrenk 2008; Cohen and Callender 2009) that a single language must be privileged as the language in which all systems competing to be the best will be expressed. Two problems have led these authors to adopt single language privileging: The first is the Trivial Systems Problem (TSP), according to which, in brief, allowing for suitably gerrymandered languages can guarantee that the "best" system will have axioms and theorems undeserving of the name "law" (see Lewis 1983 for its initial development). Language privileging provides a quick fix to the TSP as long as the privileged language is not among the suitably (and problematically) gerrymandered. The second is the Problem of Immanent Comparisons (PIC) suggested by Cohen and Callender (2009). The PIC takes it to be the case that there are only "immanent" measures for simplicity, strength, and their balance—that is, measures defined for only one language. With single language privileging, no two systems ever need to be compared when expressed in different languages, and so having to use only immanent measures is not an issue.

Though single language privileging solves these problems for the BSA and its variants, it creates new ones of its own. For one, the BSA is already often criticized for being insufficiently objective—because it is unclear that there is an objective answer to the question of what makes a system the best—and single language privileging has the potential to fuel those criticisms by requiring proponents of the BSA to say which

language gets privileged. Relativizing laws to languages (as in Schrenk 2008 and Cohen and Callender 2009) goes some way to resist such criticisms, but, as Bialek (2017) argues, relativity itself should be minimized (as much as scientific practice allows) when responding to those who employ the 'insufficiently objective' critique of the BSA. Another issue with language privileging—a version of which is suggested in a specific critique of Lewis (1983) by van Fraassen (1989), and is here newly generalized as an issue for *any* single language privinleging—is that it breaks the supposedly close connection in scientific practice between the discovery of the laws and the discovery of basic kinds.[1]

Both problems are, ultimately, overstated, and may be resolved not with single language privileging, but with the privileging of *classes* of languages. This addresses both of the issues just raised. For one, it restores the co-discovery of laws and basic kinds to the BSA by making the search for laws (via a best system competition conducted in the course of scientific practice) include a search through a class of languages for the one that yields the best system-language pair. It also helps to limit the degree to which laws may need to be relativized to language by reducing the problem of privileging a language (class) to the already present problem of choosing a measure of 'best'.

The outline of this paper is as follows. I begin, in Section 2, by laying out the PIC. In Section 3, I argue that the PIC ignores the existence of measures (illustrated by the

[1]Depending on the specific interests of the author, there has been talk of "basic kinds" (as in Cohen and Callender 2009), "fundamental kinds" (Loewer 2007), and "perfectly natural predicates" (Lewis 1983). These are progressively more restrictive ways of interpreting the predicates of a language that appear in the axioms of a best system expressed in that language. Throughout the paper I use the more general phrase "basic kinds", but nothing about that usage precludes a more restrictive reading.

Akaike Information Criterion) that, while not transcendent (since they cannot compare systems expressed in *any* two languages), are also not immanent (since they can compare systems expressed in *some* different languages). Being sensitive to the existence of such measures suggests a slightly different problem of *transcendent* measures, which may be resolved through privileging classes of languages. The problem for single language privileging of breaking the connection between the discovering laws and basic kinds is developed in Section 4, and its resolution via language-class privileging is demonstrated. In Section 5, I argue that the question of which language class to privilege is reducible to the question of which measure(s) of 'best' (simplicity, informativeness, etc.) should be used. Lastly, in Section 6, I note that the reducibility just introduced suggests a new solution to the TSP that is focused on choosing appropriate measures of 'best', with the conclusion being that none of the problems that have prompted language privileging actually require it for their resolution.

2 The Problem of Immanent Comparisons

The "Problem of Immanent Comparisons" (PIC) begins with an appeal in Cohen and Callender (2009) to a distinction in Quine between *immanent* and *transcendent* notions. Quine writes: "A notion is immanent when defined for a particular language; transcendent when directed to languages generally" (Quine 1970, p. 19). Measurements of simplicity, since they depend on the language in which a system is expressed, are taken by Cohen and Callender to be immanent in this Quinean sense. Strength, or informativeness, is similarly immanent, since it is assumed to depend on the expressive power of the language in which a system is expressed. And, to finish out the set, balance

is said to be immanent as well, since it will be a measure dependent on immanent measures of simplicity and strength. If two systems are competing to be the best and are expressed in different languages, then we would need transcendent measures of simplicity, strength, and balance, in order to implement the best system competition. But "there are too few (viz. no) transcendent measures" of simplicity, strength, and balance (Cohen and Callender 2009, p. 8). Cohen and Callender write that

> *Prima facie*, the realization that simplicity, strength, and balance are immanent rather than transcendent—what we'll call *the problem of immanent comparisons*—is a devastating blow to the [BSA and its variants]. For what counts as a law according to that view depends on what is a Best System; but the immanence of simplicity and strength undercut the possibility of intersystem comparisons, and therefore the very idea of something's being a Best System.
>
> (Cohen and Callender 2009, p. 6, emphasis in original)

The only solution to the PIC, since (supposedly) systems can only be compared when they are expressed in the same language, is to adopt single language privileging.

3 Neither Immanent nor Transcendent

The issue with the PIC is that it ignores the existence of a large middle ground of measures that are neither immanent nor transcendent. To start, let us examine the central claim of the PIC: that simplicity, strength, and balance must be immanent measures. In defense of the idea that simplicity is immanent, Cohen and Callender

(2009, p. 5) defer to Goodman (1954) by way of Loewer, who writes: "Simplicity, being partly syntactical, is sensitive to the language in which a theory is formulated" (Loewer 1996, p. 109). Loewer and Goodman are exactly right. Simplicity is language sensitive. For example, let us adopt a naive version of simplicity, $SimpC(-)$, that is measured by the number of characters it takes to express a sentence (including spaces and punctuation). Consider the following sentence.

This sentence is simple.

Its *SimpC*-simplicity is 24 characters. The same sentence in Dutch is

Deze zin is eenvoudig.

The sentence's *SimpC*-simplicity now is 22 characters. So the *SimpC*-simplicity of a sentence depends or is sensitive to the language in which the sentence is expressed. Does that language sensitivity mean that *SimpC* is immanent? It depends on what is meant by being "defined for a particular language".

SimpC is, in some sense, "defined for a particular language". Insofar as the measure gives conflicting results for a sentence expressed in different languages, it would be ill-defined if we took it to be directed at sentences irrespective of the language in which they are expressed. One way of dealing with this would be to think that we have a multitude of distinct simplicity measures: $SimpC_{\text{English}}(-)$, $SimpC_{\text{Dutch}}(-)$, and so on. But doing that disguises an important fact: each of these measures of simplicity is *the same measure*, just relativized to particular languages. Drawing our inspiration from the "package deal" of Loewer (2007)—in which the BSA holds its competition between system-language pairs (or packages)—we could just as easily deal with the language

sensitivity of *SimpC* by saying it is defined for sentence-language pairs. We don't need, then, different measures of simplicity. Just the one will do:

$$\begin{aligned} SimpC(\ulcorner\text{This sentence is simple.}\urcorner, \text{English}) &= 24 \text{ char.} \\ SimpC(\ulcorner\text{This sentence is simple.}\urcorner, \text{Dutch}) &= 22 \text{ char.} \end{aligned}$$

In this way, *SimpC* is better understood as transcendent, and not immanent, because it is, as Quine put it, "directed to languages generally".

Of course, *SimpC* can't be directed to *all* languages, since it will be undefined for any languages that don't have a written form with discrete characters. This suggest that there is an important middle ground between immanent and transcendent measures. When a measure falls in that middle, as *SimpC* seems to, I will say that it is a "moderate measure".

So which conception of *SimpC* is the right one? The "devastating blow" that immanence deals to the BSA and its variants is that it "undercut[s] the possibility of intersystem comparisons" (Cohen and Callender 2009, p. 6). In our naive example,

$$SimpC_{\text{English}}(\ulcorner\text{This sentence is simple.}\urcorner)$$

is—if *SimpC* is immanent—incomparable to

$$SimpC_{\text{Dutch}}(\ulcorner\text{This sentence is simple.}\urcorner).$$

But obviously it's not. ⌜This sentence is simple.⌝ is *SimpC*-simpler in Dutch than in English (when being *SimpC*-simpler means having a lower value of *SimpC*).

Nothing prevents a transcendent or moderate measure from taking a language as one of its arguments. Such a measure is transcendent (or moderate), but language sensitive, and, importantly, it allows for comparisons even when a variety of languages are involved. That being the case, the mere language sensitivity of simplicity, strength, and their balance is not enough to guarantee that they are immanent, nor is it enough to guarantee the incomparability of systems expressed in different languages.

In response to the existence of a measure like *SimpC*, it might be suggested that there may well be transcendent (or moderate) measures plausibly named "simplicity" (etc.), but these are not the ones relevant to the BSA; the measures that *do* appear in BSA will be immanent. It is absolutely right to question the plausibility of a measure as naive as *SimpC* having a role to play in the BSA. (I certainly do not intend to defend *SimpC* as the right measure of simplicity for the BSA.) But I do not think it is clear why we should assume that the right measures are immanent. Rather, I think that moderate measures are, if anything, the norm, and an example may be found in the selection of statistical models.

Following Forster and Sober (1994), statistical model selection has standardly been associated in philosophy with the Akaike Information Criterion (AIC):

$$AIC(M) \;=\; 2[\text{number of parameters of } M] - 2[\text{maximum log-likelihood of } M]$$

The full details of AIC are not terribly important for our purposes here; it is enough to point out that that first term is concerned with the *number of parameters* of the statistical model M. Forster and Sober note that the number of parameters "is not a merely linguistic feature" of models Forster and Sober (1994, p. 9, fn. 13). But the

number of parameters is *a* linguistic feature of a model. Since AIC can compare models with different numbers of parameters, it can—if we think of statistical models as the system-language pairs of the BSA, and AIC as central to the best system competition[2]—compare systems expressed in different languages. AIC is thus a moderate measure.

It is important to note, however, that AIC is also not a transcendent measure. Kieseppä (2001) offers a response to critics of AIC who are concerned that the measure is sensitive to changing the number of parameters of a model by changing the model's linguistic representation. The response turns on the justification of "Rule-AIC", which says to pick the model with the smallest value of AIC, on the grounds that the predictive accuracy of model M is approximately the expected value of the maximum log-likelihood of M minus the number of parameters of M. Crucially,

> the theoretical justification of using (Rule-AIC) is valid when the considered models are such that the approximation [just mentioned] is a good one.
>
> (Kieseppä 2001, p. 775)

Let M be parameterized to have either k or k' parameters. Then there are two claims that are relevant to the justification of Rule-AIC:

$$\text{predictive accuracy of } M \approx E[(\text{maximum log-likelihood of } M) - k]$$

$$\text{predictive accuracy of } M \approx E[(\text{maximum log-likelihood of } M) - k']$$

[2]To make the connection between AIC and the BSA even stronger, it it worth noting that Forster and Sober (1994) take the "number of parameters" term to be tracking the simplicity of a model.

The predictive accuracy of M is independent of the number of parameters used to express M.[3] But the right side of the approximation in each claim *does* depend on the number of parameters. In general, both of these claims will not be true. Since Rule-AIC is only justified by the truth of these approximations, it will only be applicable to whichever parameterization of M makes the approximation true. The only time when both claims are true, and thus when AIC is applicable to both parameterizations, is when the difference between $E[(\text{maximum log-likelihood of } M) - k]$ and $E[(\text{maximum log-likelihood of } M) - k']$ is negligible. Kieseppä concludes:

> This simple argument shows once and for all that the fact that the number of the parameters of a model can be changed with a reparameterisation does not in any interesting sense make the results yielded by (Rule-AIC) dependent on the linguistic representation of the considered models.
>
> (Kieseppä 2001, p. 776)

From the epistemic perspective that is Kieseppä's concern, I can find room to agree that there is no "interesting sense" in which Rule-AIC is language dependent. This is because, if we are looking to employ Rule-AIC in statistical model selection, what is available to us is a procedure to check if the given parameterization is one that can support the justification of Rule-AIC. If the justification will work, then Rule-AIC applies, and if not, not. Rule-AIC isn't language dependent "in any interesting sense" insofar as it simply doesn't apply to the problematic languages/parameterizations that undermine its justification.

[3]This is intuitively true. It is also true in the formal definition of predictive accuracy given in Kieseppä (1997) and used in this argument from Kieseppä (2001).

However, from the perspective of the BSA and the PIC, these failures of Rule-AIC *are* interesting. AIC (the measure) is not immanent, but it is also not transcendent; it is merely moderate. *Some* reparameterizations of considered models will lead to the inapplicability of Rule-AIC. If Rule-AIC was how we were deciding which system was best, the existence of these problematic reparamterizations would be, as Cohen and Callender put it, a *prima facie* devastating blow to the BSA.

Towards the end of their introducing the PIC, Cohen and Callender write that

> What is needed to solve the problem is a *transcendent* simplicity/strength/balance comparison of each axiomatization against others. The problem is not that there are too many immanent measures and nothing to choose between them, but that there are too few (viz., no) transcendent measures.
>
> (Cohen and Callender 2009, p. 8, emphasis in original)

Cohen and Callender are probably right that there are "too few (viz., no) transcendent measures". In response to this, PIC says that measuring the goodness of a system must be done with immanent measures, and so no systems expressed in different languages may be compared in the best system competition. But non-transcendence is not a guarantee of immanence. We might call the problem that remains the *problem of transcendent measures* (PTC). Measures like AIC are not immanent, but they also aren't transcendent. That non-transcendence gives rise to a degree of language sensitivity that will *sometimes* prevent us from comparing systems expressed in different languages.

In response to the PIC and the supposed immanence of measures appropriate for the BSA, Cohen and Callender (2009) proposed the Better Best Systems Analysis (BBSA),

which relativizes laws to single languages. According to the BBSA, a best system competition is run for every language L (with some restrictions on "every" that aren't especially important here) where all the competing systems are expressed in L and the theorems of the system that is the victor of the competition are the laws *relative to* L. But now it seems that we might have at our and the BSA's disposal moderate measures. In the face of the non-transcendence of these measures—that is, in the face of the PTC—the BBSA's strategy of language relativity is still a good one.[4] Our language relativity does not, however, have to involve privileging *single* languages. The alternative is to relativize to *classes* of languages constructed to ensure the applicability of the measures employed in our best system competition.

4 Discovering Laws and Kinds Together

Before saying more about what relativizing laws to classes of languages would be like in any detail, it is important to say something about why we should pursue language-class relativity over the single language relativity of the BBSA. So, why should we? The reason is that one of the great virtues of the BSA and its variants is their offering of a metaphysics for laws that parallels the search for laws that is to be found in scientific practice, and that parallel is broken by single language privileging. A feature of the

[4]Without going into excessive detail about benefits (and costs) of the BBSA's relativity strategy over competitors, I hope it is enough to note that relativizing the laws allows us to sidestep the question of which language should be privileged entirely, since, ultimately, all languages will get a turn at being privileged, and thus, effectively, none are privileged over all.

search for laws in scientific practice is that it happens in conjunction with a search for the basic kinds of the world. This feature encourages us to acknowledge the importance of language in the BSA, since the basic kinds of the world are, presumably, going to correspond with the basic kinds that appear in the language in which the laws are expressed. Thus, when Lewis first recognizes the language sensitivity of simplicity, he concludes on a celebratory note by saying that the variant of single language privileging he introduces has the virtue of "explaining" why "laws and natural properties get discovered together" (Lewis 1983, p. 368).

For Loewer's Package Deal Analysis, the idea that laws and kinds are discovered together is central to the view. Indeed, the phrase "package deal" has its roots in Lewis, who says just before the "discovered together" remark that "the scientific investigation of laws and of natural properties is a package deal" (Lewis 1983, p. 368). While Loewer ultimately endorses a version of single language privileging, it is accompanied with a rough account of how a "final theory"—i.e., a candidate system-language pair—is arrived at:

> a final theory is evaluated with respect to, among the other virtues, the extent to which it is informative and explanatory about truths of scientific interest as formulated in [the present language of science] SL or any language $SL+$ that may succeed SL in the rational development of the sciences. By 'rational development' I mean developments that are considered within the scientific community to increase the simplicity, coherence, informativeness, explanatoriness, and other scientific virtues of a theory.
>
> (Loewer 2007, p. 325)

If the practice of science parallels the Package Deal Analysis, then the processes of discovering the laws and basic kinds are one and the same.

And it seems Cohen and Callender are also on board with laws and kinds being discovered together when they offer this nice remark on the phenomenon:

> historical disputes between theorists favoring very different choices of kinds seem to us to be disputes between two different sets of laws [...] it has happened in the history of science that people have objected to particular carvings—most famously, consider the outrage inspired by Newton's category of gravity. But given the link between laws and kinds, this outrage is probably best seen as an expression of the view that another System is Best, one without the offending category. If that other system doesn't in fact fare so well in the best system competition—as in the case of the systems proposed by Newton's foes—then the predictive strength and explanatory power of a putative Best System typically will win people over to the categorization employed. While it's true that some choices of [kinds] may strike us as odd, no one would accuse science—the enterprise that gives us entropy, dark energy, and charm—as conforming to pre-theoretic intuitions about the natural kinds of the world. Yet these odd kinds are all embedded in systematizations that would produce what we would consider laws.
>
> (Cohen and Callender 2009, pp. 17–18)

With everyone in agreement, what is the problem? Language privileging, essentially, happens *before* the identification (in the BSA and its variants) or discovery (in scientific practice) of the laws. Though Cohen and Callender will not "accuse science" of

"conforming to pre-theoretic intuitions about the natural kinds of the world", that is exactly what the BBSA (and any other single language privileging variant of the BSA) does when it privileges sets of kinds prior to a best system competition. Furthermore, PIC makes it such that "the predictive strength and explanatory power of a putative Best System" cannot "win people over to the categorization employed" because comparing two putative Best Systems expressed in different languages (with different "categorizations") is supposed to be impossible.[5]

Relativizing to classes of languages solves this problem. Scientists are able to approach the discovery of laws and kinds with pre-theoretic intuitions about how to systematize the world, the language to use when doing that, and the best system competition. As we will see below, the intuitions regarding language and the best system competition will locate them in a particular language class. Scientists will move away from their intuitions about language (and systematizing) when, much as Loewer describes above, there are languages in the relevant language class that may be paired with systems to yield a system-language pair that is scored better by the best system competition than the pre-theoretic system-language pair.[6]

[5]At least, it is impossible according to PIC for the BSA and its variants. If it *is* possible for scientists, then it is wholly unclear why it would be impossible for the BSA.

[6]This movement is only metaphorical for the BSA, where all the possibilities are considered and judged simultaneously. It is helpful, though, to think in the more methodical terms—of considering particular transitions from one system-language pair to another, the benefits that they might bring, and then adopting them or not—because that is what will happen in actual scientific practice.

5 Limiting Language Relativity

Let us begin addressing how language-class relativity can work by looking in more detail at the single language relativity of the BBSA. In the BBSA, there are the fundamental kinds K_{fund}. The set of all kinds $\mathcal{K}$ is the set including K_{fund} closed with respect to supervenience relations—that is, $\mathcal{K}$ includes every kind that can be defined as supervening on the arrangement of the K_{fund} kinds in the actual world. A language L is determined by the set of kinds for which it has basic predicates, and there is a language L_i for every $K_i \subseteq \mathcal{K}$. For any two languages L and L', the supervenience relations between the kinds of the languages and K_{fund} can be thought of as schemes for *translation* between L and L'. The set of all languages $\mathcal{L}_{\text{all}}$ can be thought of as the set of languages that includes L_{fund} closed with respect to all translations. A class of languages $\mathcal{L}_i$ is a set of languages including L_{fund} closed with respect to some acceptable (all, in the case of $\mathcal{L}_{\text{all}}$) translations.

To illustrate, let us consider a 'coin flip' world. Such a world is a string of `H`s and `T`s, which we will assume are the only two fundamental kinds. Another set of kinds might be $K_{\text{ex}} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$, where the translation that gets us to the corresponding language L_{ex} from L_{fund} maps the pairs `HH`, `HT`, `TH`, and `TT`, to `a` through `d`, respectively. An example of a class of languages that includes L_{ex} could be $\mathcal{L}_{\text{n-tuple}}$: Let an acceptable translation for $\mathcal{L}_{n\text{-tuple}}$ be one that, for a given n takes the set of all n-tuples of `H` and `T`, and maps them to a set of kinds $K_n = \{k_{n,1}, k_{n,2}, ...k_{n,2^n}\}$. L_{fund}, then, is just L_1. When `a` through `d` are $k_{2,1}$ through $k_{2,4}$, our K_{ex} and L_{ex} are precisely K_2 and L_2. All, and only, the languages that may be formed through this procedure will be members of the class $\mathcal{L}_{\text{n-tuple}}$.

A language-class relative variant of the BSA will run a best system competition for

every class of languages $\mathcal{L}_i$. Then $\mathcal{S}$ is the set of all systematizations of the world, the set of all competing system-language pairs for the $\mathcal{L}_i$-relative best system competition is given by $\mathcal{S} \times \mathcal{L}_i$.

We can apply this conception of language-class relativity to our other running example of statistical model selection with AIC. Recall that *some* reparameterizations of statistical models would prove problematic for the use of AIC. To reparameterize a model is akin to translating it from one language to another. We can understand, then, the problem of language sensitivity for AIC as being related to some set of problematic translations. If we subtract these problematic translations from the set of all translations, then we have a set of acceptable translations which defines a class of languages that we can call $\mathcal{L}_{AIC}$. $\mathcal{L}_{AIC}$ is precisely the set of all languages such that a system expressed in any one of them will be comparable to a system expressed in any other using AIC. As long as the moderate measures used in the best system competition have clearly problematic and/or acceptable translations associated with them, then the class of languages that may be used to express competing systems will be determined by the measures used in the best system competition.

This will have one of two effects on the extent to which the BSA must be relativized to classes of languages, but before going into those details it will be helpful to characterize "competition relativity". Competition relativity should be understood in much the same way that language relativity is understood. The competition of the BSA is the thing that takes system-language pairs as its inputs, and outputs a best pair from which we can read off the laws. The competition decides what system-language pair is best by considering how well they measure up with respect to some collection of theoretical virtues (like simplicity and informativeness) and the actual world. Much as

we might worry about what language to privilege, and side-step that problem by relativizing laws to languages so that every language takes a turn as the privileged one, we might also worry about which competition, or which set of theoretical virtues, to privilege. Competition relativity sidesteps the problem of which collection of theoretical virtues to use (and weighting between them, and means of measuring them, etc.) by relativizing laws to every way of formulating a best system competition.[7]

So, either the BSA will be committed to competition relativity or not. Suppose that it is not. For convenience, suppose further that Rule-AIC is all that there is to the best system competition. In that case, the BSA will always be run using the $\mathcal{L}_{AIC}$ class of languages. Language-class relativity is not required since there is only one language class that will ever be relevant to the BSA—namely $\mathcal{L}_{AIC}$, as determined by the best system competition. Now suppose that there is competition relativity. A different best system competition must be run for every competition function C_i in the set of all possible competition functions $\mathcal{C}$. In principle we will need to run best systems competitions for every pair in $\mathcal{C} \times \mathbb{L}$, where $\mathbb{L}$ is the set of all language classes. Let $\mathcal{L}_j$ be the class of languages constructed according to the translations that are acceptable for the measures that comprise C_i when $i = j$. In practice, however, it will only make sense to run a competition once for each $C_i \in \mathcal{C}$, since the pairs $C_i, \mathcal{L}_j$ will be unproblematic only when $i = j$. Language-class relativity in this situation will be redundant with competition relativity. We also have it that, in either case (of needing competition relativity or not), single language relativity remains unnecessary for all the same reasons that recommended language-class relativity.

[7]See Bialek (2017) for an extended discussion of competition relativity and the possibility of its inclusion in the BSA.

6 The Trivial Systems Problem

The redundancy of any sort of language privileging relativity with competition relativity offers an interesting solution to the Trivial Systems Problem (TSP) that initiated the trend of single language privileging.

Recall that the TSP is concerned with the possibility of suitably gerrymandered languages that can guarantee that the "best" system will have axioms and theorems undeserving of the name "law". In the introduction fo the problem, Lewis imagines a system S and predicate F "that applies to all and only things at worlds where S holds" (Lewis 1983, p. 367). The system S, then, maybe be expressed by the single axiom $\forall xFx$, simultaneously achieving incredible informativeness—because of the specific applicability of F—and incredible simplicity—because, Lewis assumes, '$\forall xFx$' is about as simple as a system could be. So S will be the best system despite a variety of reasons why it shouldn't be, the foremost of which are that: (1) $\forall xFx$ will be a law unlike any we would expect to find, (2) F would be a basic kind unlike any we would expect to find, and (3) every regularity of the world is a theorem of $\forall xFx$, so there would be no distinction between accidental and lawful regularities.

The problem is solved as long as we can avoid languages that include problematic predicates like F. Single language privileging solves this problem as long as the privileged language does not include the (or any) problematic predicate(s). Language-class privileging likewise solves the problem as long as no language in the class includes the (or any) problematic predicate(s). That alone might be enough said, but the redundancy of language-class choice on competition choice offers a more nuanced solution: The best system competition could be chosen such that the corresponding class

of languages does not include F or any similarly problematic predicates. But it could also be chosen such that F and its ilk are certain to not be the best. Lewis assumes with no discussion that $\forall xFx$ is an incredibly informative and simple system, but, even if that is true for the measures/competition, it need not be true for every competition. If there is competition relativity, then there may be competitions for which a trivial system like $\forall xFx$ is the victor, but for the same reasons that such a system is problematic, scientists will simply be uninterested in the laws relative to those competitions.[8] If there isn't competition relativity, it seems unlikely that science would unequivocally endorse a competition that yields a trivial system (or, if it does, then we would need to take a step back and seriously reconsider our aversion to such a system).

In the end, there is no apparent need for any language privileging or relativity in the BSA.[9] Its role in solving the problems of immanent (or transcendent) comparisons and trivial systems will be unnecessary (if a single moderate best system competition can be identified) or redundant with competition relativity.

[8]In much the same way that Cohen and Callender (2009) allow for there to be uninteresting sets of laws determined relative to languages that include F-like predicates.

[9]The problems discussed is not the only reason one might want to adopt language relativity in the BSA. It should also be noted that one of the virtues of the BBSA's single language relativity is that it allows the view to accommodate an egalitarian conception of special science laws. Language relativity, however, is not the only way of getting special science laws out of the BSA. This is an important issue to which the discussion in this paper is relevant, but a proper exploration of it warrants a more focused and extended treatment.

References

Bialek, M. (2017). Interest relativism in the best system analysis of laws. *Synthese 194*(12), 4643–4655.

Cohen, J. and C. Callender (2009). A better best system account of lawhood. *Philosophical Studies 145*(1), 1–34.

Forster, M. and E. Sober (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science 45*(1), 1–35.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Kieseppä, I. (1997). Akaike information criterion, curve-fitting, and the philosophical problem of simplicity. *The British journal for the philosophy of science 48*(1), 21–48.

Kieseppä, I. (2001). Statistical model selection criteria and the philosophical problem of underdetermination. *The British journal for the philosophy of science 52*(4), 761–794.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Loewer, B. (1996). Humean supervenience. *Philosophical Topics 24*(1), 101–127.

Loewer, B. (2007). Laws and natural properties. *Philosophical Topics 35*(1/2), 313–328.

Quine, W. V. O. (1970). *Philosophy of logic*. Harvard University Press.

Schrenk, M. (2008). A theory for special science laws. In S. W. H. Bohse, K. Dreimann (Ed.), *Selected Papers Contributed to the Sections of GAP.6*, pp. 121–131. Paderborn: Mentis.

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -144-

Explaining Scientific Collaboration: a General Functional Account

Thomas Boyer-Kassem* and Cyrille Imbert†

October, 2018

Abstract

For two centuries, collaborative research has become increasingly widespread. Various explanations of this trend have been proposed. Here, we offer a novel functional explanation of it. It differs from accounts like that of Wray (2002) by the precise socio-epistemic mechanism that grounds the beneficialness of collaboration. Boyer-Kassem and Imbert (2015) show how minor differences in the step-efficiency of collaborative groups can make them much more successful in particular configurations. We investigate this model further, derive robust social patterns concerning the general successfulness of collaborative groups, and argue that these patterns can be used to defend a general functional account.

*MAPP (EA 2626), Univ. Poitiers, France. thomas.boyer.kassem@univ-poitiers.fr
†CNRS, Archives Poincaré, France. cyrille.imbert@univ-lorraine.fr

1 Introduction

For two centuries, co-authoring papers has become increasingly widespread in academia (Price, 1963, Beaver and Rosen, 1979), especially in the last few decades. Since the 1950s, the percentage of co-authored papers has grown at a common rhythm for science and engineering, social sciences, and patents; the mean size of collaborative teams has also increased, and even more so in science and engineering. No such increase is visible for the art and humanities (Wuchty et alii, 2007).

Various explanations of this collaborative trend have been proposed: for example, it may be caused by scientific specialization, it may increase the productivity or reliability of researchers, or be promoted by the rules of credit attribution. Here, we aim at offering a new functional explanation of this trend by showing that collaboration exists because it increases the successfulness of scientists. The present explanation differs from accounts like that of Wray (2002) by the social and epistemic mechanism that grounds the beneficialness of collaboration. We analyze further an existing model that shows how minor differences in the step-efficiency of collaborative groups at passing the steps of a project can make them much more successful in particular configurations (Boyer-Kassem and Imbert, 2015) and show how it can be used to build a general and robust functional explanation of collaboration.

We introduce the model in section 2. After presenting functional explanations (section 3), we show how the model can be used to derive robust social patterns of the successfulness of collaborative groups (section 4), and argue that these patterns can refine and strengthen functional explanations of collaboration like the one defended by Wray (sections 5 and 6).

2 Boyer-Kassem and Imbert's Model: Main Results and Explanatory Lacunas

Boyer-Kassem and Imbert (2015) investigate a model in which n agents struggle over the completion of a research project composed of l sequential steps. At each time interval, agents have independent probabilities p of passing a step. When an agent reaches the end of the project, she wins all the scientific credit and the race stops (this is the priority rule). Agents can organize themselves into collaborative groups for the whole project, meaning that they only share information, i.e. step discoveries — clearly, there are more favorable hypotheses associated with collaborating, like having new ideas or double-checking (see below). Within a group, agents make progress together, and equally share final rewards. Thus, a group of k agents (hereafter k-group) passes a step with probability $p_g(k,p) = 1-(1-p)^k$. In forthcoming illustra-

tions, the value of l is set to 10 and that of p to 0.5, which is not particularly favorable for groups (ibidem, 674). If collaboration is beneficial with these hypotheses, it will be even more so with more favorable or realistic ones. A community of n agents (hereafter, n-community) can be organized in various k-groups. For example, a 3-community can correspond to configurations (1-1-1), (2-1) or (3). The individual successfulness of an agent in a k-group in a particular configuration is defined as the average individual reward divided by time. It has been obtained for all configurations up to $n = 10$, on millions of runs.

Note that this model is not aimed at quantifying the actual successfulness of collaborative agents, but at analyzing the differential successfulness of agents depending on their collaborative behavior. The main finding is that minor differences in the efficiency at passing steps can be much amplified and that, even with not-so-favorable hypotheses, collaboration can be extremely beneficial for scientists. For example, in a (5-4) (resp. (2-1)) configuration, whereas the difference in step efficiency between the 5 (resp. 2) and the 4-group (resp. 1-group) is 3% (resp. 50%), the difference in individual successfulness is 25% (resp. 700%). The scope of these results actually goes beyond the initial hypotheses in terms of information sharing. Formally speaking, the model is a race between (collective) agents i with probabilities p_i of passing steps. *Whatever the origin* of the differences in p_i, they are greatly amplified by the sequential race. In other words, any factor, whether epistemic or not, that implies an increase in p_i of a k-group (e.g. if a collaborator is an expert concerning specific steps, if increased resources improve step-efficiency, etc.) makes this group as successful as a larger group — hence the generality of this mechanism.

Still, these results do not explain scientific collaboration by themselves. First, collaboration is beneficial for particular k-groups in particular configurations only: a 2-group is very successful in configuration (2-1-1-1-1) but not in (7-2). Thus, the model mostly provides possibility results about what can be the case in certain configurations. Second, the explanandum is a general social feature of modern science, not some collaborative behavior in some particular case, so the explanans must also involve general statements about the link between collaboration and beneficialness. Then, if the model presents generic social mechanisms with explanatory import, one needs to describe at a general level the effects of these mechanisms and provide some general, invariant pattern between collaboration and beneficialness. This is what we do in section 4. A final serious worry is that the beneficialness of a state by no means explains why it exists, nor perseveres in being. A link needs to be made between the beneficialness of collaboration and its existence over time. We suggest that this connection can be accounted for functionally.

3 Functional Explanations and Collaboration

We review in this section how functional explanations work and how they can be used in the present case. We follow Wray's choice to use Kincaid's account because it is simple, widely accepted, and that nothing substantial hinges on this choice. Functional explanations explain the existence of a feature by one of its effects, usually its usefulness or beneficialness. As such, they can be sloppy and badly flawed. The usefulness of the nose to carry glasses does not explain that humans have one. Nevertheless, if stringent conditions are met, it is usually considered that functional explanations can be satisfactory, typically within biology. Even Elster, who otherwise favors methodological individualism, agrees that functional explanations can be acceptable in the social science (Elster, 1983). According to Kincaid (1996, 105-114), P is functionally explained by E, i.e. P exists "in order to promote <effect E>" if:

(1) P causes E,

(2) P persists because it causes E,

(3) P is causally prior to E.

Then, a functional explanation of collaboration should have the following form:

(1c) Scientists' collaborative behavior causes the increase of their individual successfulness.

(2c) Scientists' collaborative behavior persists (or develops) because it causes a higher individual successfulness.

(3c) Collaborative behavior is causally prior to this increased individual successfulness that is rooted in collaborative behavior.

We agree with Wray (2002, 161) that it is implausible to consider that the high successfulness of scientists is the initial cause of collaboration since many scientists have been successful (and continue to be in some fields) without collaborating. In the same time, there can be various contingent reasons why some researchers have decided to engage in some collaboration. So, what calls for an explanation is the fact that collaboration is widespread and persistent, not its occasional existence.

4 Collaboration Causes Successfulness

We now argue that the above model provides strong evidence in favor of (1c). To explain the general collaborative patterns described above, the causal

relation between collaboration and successfulness needs to be general and robust. Hence, one needs to go beyond the description of the beneficialness of collaboration in particular situations. A first route is to find general results about when it is beneficial for individuals to collaborate, such as the following theorem (see the appendix for the proof).

Theorem. When m groups of equal size k merge, the individual successfulness of agents increases.

In other words, as soon as several k-groups of the same size exist, they would improve the individual successfulness of their members by merging. A corollary is that single individuals always have interest in collaborating. However, this theorem only covers a small subset of possible configurations, and cannot provide a general vindication for the causality claim (1c). Further, agents might only use it if they are aware of it and are in a position to identify groups of equal-size competitors, which cannot be assumed in general.

To overcome these difficulties, we now assess agents' successfulness irrespective of what they know about other competitors: we consider the average successfulness of k-groups over all possible configurations for each community size. For example, we average the individual successfulness of 4-groups in configurations (4-1-1-1); (4-2-1) and (4-3)[1]. In order to study the robustness of the causal relation between collaboration and successfulness, we investigate in the next paragraphs how much collaborating remains beneficial under variations of key parameters of the competition context.

Successfulness and community size. Figure 1 shows the average successfulness within k-groups for communities of various sizes. First, the successfulness of loners brutally collapses and is much lower than that of other k-groups as soon as $n > 2$. This confirms that except when nobody collaborates, or in very small communities, loners are outraced. Second, for all group sizes, individual successfulness decreases for larger communities, as can be expected when the number of competing groups and their size increases. Nevertheless, the successfulness of k-groups remains high and stable up to some community size s larger than k till they are eventually outperformed by larger groups or till growing bigger would mean over-collaborating (see (Boyer-Kassem and Imbert, 2015, 679-80) for an analysis of over-collaboration in large groups). Third, the larger the groups are, the longer and flatter this initial plate of successfulness is and the less steep the decrease in successfulness is. Fourth,

[1]There is no clear rationale about how to weigh configurations. From a combinatorial viewpoint, configuration (1,1,1,1,1,1,1) has one realization and (3,2,1,1) several ones. But from an empirical viewpoint, when scientists hardly collaborate, configuration (1,1,1,1,1,1,1) is usual and (3,2,1,1) extremely rare. We have privileged simplicity and chosen to give equal weight to all configurations.

when n is much larger than k, the successfulness of k-groups increases with k. However, this increase is a moderate one and small groups still do reasonably well, which is somewhat unexpected, given the general amplification effect — but see the analysis of figure 3 below for more refined analyses. Typically, in 10-communities, 2-groups do badly but remain somewhat viable since their average successfulness remains between 1/3 to 1/2 of that of 3 or 4-groups. Overall, not collaborating is in general not a viable strategy. Collaborating moderately ($k = 2$ or 3) can be very rewarding when there are few competitors (e.g. in small research communities, or on ground-breaking questions that are only known to a handful of scientits). Smalls groups remain viable but tend to be outraced when communities become significantly larger (typically, concerning questions belonging to normal science that many researchers are likely to tackle). Thus, moderately collaborating is a viable but more risky strategy when uncertainty prevails about the number and size of competing groups. Finally, while large collaborative groups rarely get exceptionally high gains, they are extremely safe, with moderate differences in successfulness between them or when the community size increases.

Successfulness and group size. Figure 2 shows the variation of individual successfulness with group size for various community sizes. First, for $n > 2$, the successfulness curve has a one-peaked (discrete) form, the maximum of which grows with the community size. Second, these one-peaked curves are not symmetric: the increase in successfulness is steep (but less so for larger groups), the decrease is gradual (idem). Large groups predate resources so groups need to grow big quickly to get some share and because returns can be increasing (Boyer-Kassem and Imbert 2015, 678), the increase in successfulness is steep. The decrease after the peak is slow because large group are hard to predate but over-collaborating can become suboptimal when the increase in gain by predation no longer makes up for the need to share between more people). These results are not trivial because at the configuration level, the successfulness of groups is contextual. They are important, too. A one-peaked profile is usually *assumed* in the literature about coalitions. Here, it emerges from a micro-model, and gets its justification from it. Overall, these patterns show again that agents have a large incentive to collaborate substantially, whatever the competing environment.

Successfulness in more or less collaborative communities. Figure 3 finally shows how the successfulness of k-group members varies with the degree of collaboration in their competition environment.[2] Here again, what matters

[2]Here, the degree of collaboration in each configuration is assessed by computing the average size of k-groups. For each k, we then compute the average successfulness of a member of a k-group over configurations having a degree of collaboration within intervals $[1, 1.5]$ (represented at coordinate "1.25" on the x-axis), $[1.25, 1.75]$, $[1.5, 2]$ ... $[3.5, 4]$. We

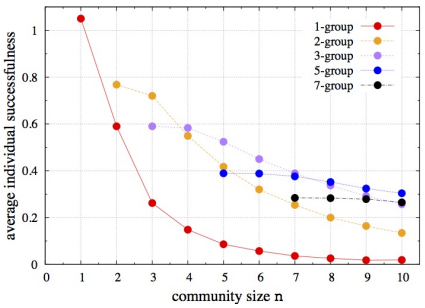


Figure 1: Variation of individual successfulness with community size.

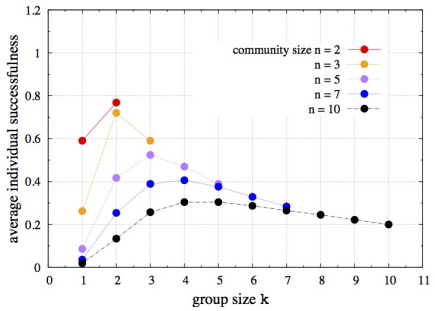


Figure 2: Variation of individual successfulness with the size of groups.

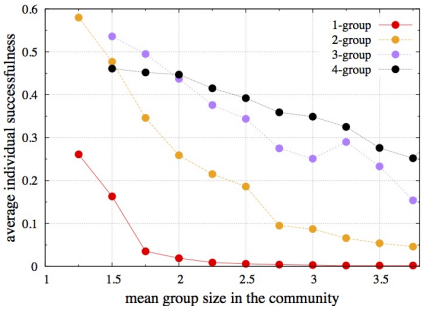


Figure 3: Variation of successfulness with the degree of collaboration in communities.

is less the exact value of the successfulness than the differential successfulness between more or less collaborating individuals. The graph confirms that successfulness depends less on the absolute size of groups than on how much they collaborate in comparison with their competitors. Scientists who collaborate more than average are very successful; those who collaborate as their peers do reasonably well; those that collaborate less than average are outraced by a large margin. This general result is not unexpected given all the above results, but the graph highlights that success for intensively collaborating scientists, and underachievement for under-collaborators can be very large. This is an important finding because if, as we shall see, successful scientists pass over their collaborative habits more than their peers, then the feedback loop provides a mechanism that favors the *increase* of the degree of collaboration by promoting those that collaborate more than others.

have chosen overlapping intervals to smoothen results. The average is computed up to communities of size 10.

Partial conclusion. Overall, the results show that — everything else being equal — collaborating a lot entails successfulness. This relation is robust under changes in the size of communities or in the exact size of groups. Further, those who collaborate more than average are much more successful. Collaborating too much is not a significant problem, under-collaborating is. So, collaborating a lot is a safe working habit, especially in the absence of information about the size and structure of the competing community. In light of this evidence, (1c) seems adequately supported.

5 Collaborative Practices Develop Because of the Success of Collaborative Scientists

We have so far argued that collaborative scientists, especially when they collaborate more than others, are more successful. We now need to argue that, because of this differential successfulness, collaborative habits persist and possibly develop in scientific communities (2c). A wide variety of social mechanisms across scientific contexts can contribute to this feedback loop. Accordingly, we shall be content with giving various evidence that strongly suggests that this link is a likely one.

Transmission. Knowing how and when to collaborate is not straightforward. Like other know-how skills, it can be developed by exercising it with people who already possess the relevant procedural knowledge. In this case, people who already collaborate can endorse this role of cultural transmission for colleagues and above all students (Thagard, 2006). Working with students is an efficient way to train them as scientists (Thagard, 1997, 248—50), so scientists have incentives to enroll students in their collaborative groups. Then, the cultural transmission of collaborative practice does not require any particular effort on top of that. The very circumstances that make collaboration possible and beneficial also make its transmission easier: when a research project can be divided into well-defined tasks, the solutions of which can be publicly assessed and shared, it is easier to enroll other people and thereby transmit collaborative skills to them (ibidem). Thus, collaborative habits can be passed over and need not be reinvented by newcomers.

Transmission opportunities. We now argue that collaborative scientists, because they are more successful, will more often be in a position to transmit their collaborative habits and that the collaboration rate will therefore increase. Within applied science, in which collaboration is also widespread (Wuchty, 2007), research projects are usually directed at finding profitable applications, which can be patented. Thus, fund providers are directly and strongly interested in hiring and providing resource to successful scientists,

who develop such applications. Within pure science, the connection is less straightforward. But because scientific success is the official goal of science, successful scientists can be expected to stand better chances to get good positions and grants, develop research programs, and pass over their collaborative habits.

Note that it is merely needed that the function between the pragmatic rewards of scientists and their success is on average increasing. This remains compatible with the fact that *some* epistemically successful scientists get little resource and *some* unsuccessful scientists get a lot — which seems to be the case. Actually, non-epistemic factors may even tend to over-credit successful scientists, and in particular collaborative ones. First, individual successfulness has been assessed in the model with a conservative estimate. It seems that an agent's publication within a k-group is actually more appreciated than just $1/k$ of a single-authored publication. For instance, a large French research institution in medicine officially weighs the citations of a paper with "a factor 1 for first or last author, 0.5 for second or next to last, and 0.25 for all others" (Inserm 2005). Also, a publication within a 10-group will generally be more visible than one single-authored publication, since more people can promote or publicize collective publications and research topics. Second, sociology of science seems to indicate that scientific credit tends to accrue to a subset of scientists who are perceived as extremely successful — this is the Matthew effect (Merton, 1968). Then, to the extent that access to resources increases with scientific credit, successful collaborative scientists can be expected to benefit from this effect and transmit more their working habits. The concentration of credit and resource may further stimulate collaborative behavior with these fortunate scientists.

Other types of mechanisms may contribute to this process, like conscious ones. So far, agents have only been supposed to follow their working habits and sometimes transmit them. But supplementary intentional or imitative processes may also feed this dynamics [3]. Once winners of the scientific race publish co-authored articles, it becomes easy for others to see that successful scientists are highly collaborative ones. (For instance, if agents of a 3-group are 4 times more successful than a single agent, this means that their groups publishes 12 more articles than this agent). Accordingly, the belief that collaborating is beneficial can be acquired as collaborating becomes usual. Furthermore, resources may accrue to scientific institutions that host individually successful scientists, and indirectly to these scientists. Agents in the model can be reinterpreted as teams or collective entities which decide to share results or to combine their expertise to produce collective articles. Then, these institutions

[3]Kincaid mentions that "complex combinations of intentional action, unintended consequences of intentional action, and differential survival of social practices might likewise make these conditions [(1)–(3) in our Section 3] true" (Kincaid 1996, 112).

and their members will be more successful, may attract resource, and will keep developing and transmitting their working habits.

In light of the above discussion, we believe that the causal connection between the success of collaborative scientists and the persistence and development of collaborative practices is highly plausible.

6 Discussion

Good functional explanations should be unambiguous about when the causal mechanisms that they rely on are efficient. In the present case, the following conditions can be emphasized.

First, conditions for the application of the priority rule should be met. In particular, (i) it should be possible to single out problems and to state uncontroversially when they are solved. Second, for the model to apply, (ii) scientific problems should be dividable into subtasks, and (iii) the solutions of these subtasks should be communicable. Finally, the model assumes that (iv) the completion of these subtasks should be sequential, but our conclusions still hold if this condition is relaxed. Indeed, if some subtasks can be tackled in parallel then the project can be completed even more quickly by different agents of a group, and collaboration is even more successful. Conditions (i)-(iii) are somewhat met in the formal and empirical sciences, less so in the social science, and almost not in the humanities. For example, as noted by Thagard (1997, 249), the humanities do not obviously lend themselves to the division of labor and to teacher/apprentice collaborations. Similarly, the importance of interpretative methods and the coexistence of incompatible traditions may prevent consensus on the nature of significant problems and what counts as a solution. This may account for the differences concerning collaborative patterns in these fields.

As mentioned above, different causal pathways may connect the successfulness of collaborative scientists to the persistence and development of collaborative practices. Thus, conditions for the fulfillment of claim (2c) cannot be uniquely specified. But several points are worth mentioning. First, the activity of epistemically successful scientists should be favored by scientific institutions. This can be the case if it is agreed that scientific success, in the form of publications or patents, is valued and promoted. Concerning scientific results that lead to patents, applications and financial gains, this condition is met when public or private funders value such outputs. Concerning pure scientific results, this means that there should be a wide agreement about which results are scientifically good and significant, and there should exist common and accessible publication venues, the value of which is consensual. Again, these conditions are approximately met in the formal and empirical sciences, less so in the social science and, almost not in the humanities in which scholars do not share paradigms, methods or norms about what is scientifically sound

and significant, and cultural and linguistic barriers can restrain the existence of unified communities and common publication venues. Second, in contexts in which researchers and projects are regularly evaluated, especially by agents or institutions who are not in a position to asses the scientific value of their work, the existence of a common standard of success in terms of publications (through simple and calibrated publication indicators) may even more favor researchers who are successful, and therefore the development of collaboration. Finally, when resources are crucial to carry out or facilitate research, snowball effects can favor even more successful scientists, and in particular collaborative ones. This resource accessibility condition, which is central in Wray's explanation, is not in ours. But we agree that in such cases, the functional mechanisms that we describe will be even stronger. In this sense, our account encompasses Wray's. This condition about resources may be another reason for the difference in collaborative behavior between the formal or empirical sciences, the social sciences and the humanities.

7 Conclusion

We have argued that collaborating a lot is overall a safe and success-conducing practice. This conclusion is robust for various sizes of groups, communities and degrees of collaboration; everything being equal, those who collaborate more than average do better. Then, to the extent that the successfulness of researchers gives them more opportunities to transmit their research habits, the development of collaborative practices in communities can be functionally explained. We have further emphasized that the conditions for this functional pattern to work are specifically met in the scientific fields in which collaboration is well-developed. Accordingly, it seems reasonable to consider that this functional mechanism is an important element of the explanation of the development of collaboration in modern science.

The explanation of collaboration is probably a multi-factorial issue. Nevertheless, an asset of our general functional explanation is that it highlights the unexpected force of beneficial aspects of collaborative activities and suggests important roles for contextual factors that are associated with the rise of collaboration. As such, it is general and unifying. For instance, the competition model shows how the division of scientific labor, the use of specialized experts (Muldoon 2017), or the increased reliability of collaborative teams (Fallis 2006, 200) can increase the probability that groups pass research steps and have amplified effects in terms of successfulness. Similarly, factors like the need to access resources to carry out or facilitate research can create a snowball effect that favors epistemically successful (collaborative) researchers (Wray 2002). And factors like the globalization of research or professionalization (Beaver, 1979) can be seen as conditions favoring the application of the priority rule

and scientific competition.

Finally, while nothing in the model provides an internal limit to the growth of collaboration, one can note that there is a wealth of reasons why collaborating groups cannot develop forever. For example, communities are limited in size, spatially distributed, and collaboration is all the more costly as groups are large. The model could be easily modified to integrate factors that limit the success and development of collaboration.

8 Appendix: Proof of the Theorem

Consider first the simple case where the m k-groups don't have other competitors. By symmetry, all groups have the same probability $1/m$ to win the race and get the reward — call this reward r. So, the individual expected reward is $r/(km)$. Suppose now the groups merge and all km agents collaborate. Each of them will receive the same reward, so their expected individual rewards are $r/(km)$ too. However, what matters in the model is not the expected reward, but the successfulness, which is this quantity divided by time. Because within a collaboration agents share all the steps they pass, the larger km-group will be at least as quick, and sometimes more, than all k-groups — more precisely: for a given drawing of all random variables corresponding to attempts to pass the steps, for all agents and temporal intervals, the km-group will move at least as quickly as all k-groups. So the individual successfulness is at least as high when identical groups merge.

Consider now the case where there are other competitors than the m groups. For a given drawing of all random variables, either the winner is one of the m groups, or another competitor. In the former case, the above reasoning can be made again, and the same conclusion holds. In the latter case, there is nothing to lose, and because the km-group is sometimes quicker than the m k-groups, there can be additional cases where it outcompetes the other competitors; then, the individual successfulness increases with the merging. QED.

9 References

Beaver, Donald deB. and Rosen, Richard (1979) "Studies in Scientific Collaboration: Part III", *Scientometrics*, 1(3): 231-245.

Boyer-Kassem, Thomas, and Cyrille Imbert (2015), "Scientific Collaboration: Do Two Heads Need to Be More than Twice Better than One?" Philosophy of Science 82 (4): 667–88.

Elster, Jon (1983), *Explaining Technical Change: A Case Study in the Philosophy of Science*, Studies in Rationality and Social Change, New York: Cambridge University Press.

Fallis, Don (2006), "The Epistemic Costs and Benefits of Collaboration", *Southern Journal of Philosophy* 44 S: 197–208.

INSERM (2005), "Les indicateurs bibliométriques à l'INSERM", `https://www.eva2.inserm.fr/EVA/jsp/Bibliometrie/Doc/Indicateurs/Indicateurs_bibliometriqu Inserm.pdf`

Kincaid, Harold (1996), *Philosophical Foundations of the Social Sciences*, Cambridge University Press.

Merton, Robert K. (1968), "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered", *Science*, 159 (3810): 56–63.

Muldoon, Ryan (2017), "Diversity, Rationality, and the Division of Cognitive Labor", in Boyer-Kassem, T., Mayo-Wilson, C. and Weisberg, M. (eds.), *Scientific Collaboration and Collective Knowledge*, New York: Oxford University Press.

Price, Derek John de Solla (1963), *Little Science, Big Science*, New York, Columbia University Press.

Thagard, Paul (1997), "Collaborative Knowledge", *Nous* 31(2): 242—261.

— (2006), "How to Collaborate: Procedural Knowledge in the Cooperative Development of Science", *The Southern Journal of Philosophy*, XLIV: 177—196.

Wray, K. Brad (2002), "The Epistemic Significance of Collaborative Research", *Philosophy of Science* 69 (1): 150-168.

Wuchty, Stefan, Jones, Benjamin F. and Uzzi, Brian (2007), "The Increasing Dominance of Teams in Production of Knowledge", *Science* 316(5827): 1036-1039.

Individuating Genes as Types or Individuals: Philosophical Implications on Individuality, Kinds, and Gene Concepts

Ruey-Lin Chen
Department of Philosophy
National Chung Cheng University
This paper will be presented at PSA 2018 meeting at Seattle in November

Abstract

"What is a gene?" is an important philosophical question that has been asked over and over. This paper approaches this question by understanding it as the individuation problem of genes, because it implies the problem of identifying genes and identifying a gene presupposes individuating the gene. I argue that there are at least two levels of the individuation of genes. The transgenic technique can individuate "a gene" as an individual while the technique of gene mapping in classical genetics can only individuate "a gene" as a type or a kind. The two levels of individuation involve different techniques, different objects that are individuated, and different references of the term "gene". Based on the two levels of individuation, I discuss important philosophical implications including the relationship between individuality and individuation and that between individuals and kinds in experimental contexts. I also suggest a new gene conception, calling it "the transgenic conception of the gene."

Keywords: gene concept, individuality, individuation, experiment, classical genetics, transgenic technique

1. Introduction: what is a gene and why individuation matters

"What is a gene?" and its related questions have been asked over and over by philosophers, historians, and scientists of biology (Beurton, Falk, and Rheinberger 2000; Carlson 1991; Falk 1986, 2010; Gerstein et al. 2007; Griffths and Stotz 2006, 2013; Kitcher 1982, 1992; Pearson 2006; Stotz and Griffiths 2004; Snyder and Gerstein 2003; Waters 1994, 2007). Those questions are frequently embedded in discussions about the definition of the term "gene" and the gene concept. As a consequence, the phrase "a gene" in this question usually refers to a type of gene. However, should we use "a gene" to refer to an individual gene, i.e., a gene token? Could it in fact be this?

The question "what is a gene" explicitly implies the problem of identifying genes, and identifying a gene presupposes individuating the gene. In what ways are genes individuated and how do scientists individuate them? I call this *the individuation problem of genes*. This paper shall approach the problem from three different but related perspectives.

From the epistemic perspective, a concept of the gene provides at least a working definition, which by nature is a hypothesis, for scientific research. Any hypothesis of the gene may be in error and may be confirmed only by experimentally individuating particular tokens of some gene. From the semantic perspective, according to a Fregean philosophy of language, the concept of reference usually serves for proper names that refer to individuals or particulars. We may extend the concept of reference to general terms (e. g., "humankind" or "gene kind") for the case in which some token of a kind is presented, and so we use a general term to refer to the kind. This means that at least some token of a kind has to be individuated. This semantic perspective presupposes an ontological perspective: the existence of a kind should be presented or demonstrated by the existence of at least a token of the kind. In the case of the gene, the ontological requirement means that we have to individuate a token of some gene kind. All three perspectives indicate the key status of individuation for answering the question of what a gene is.

According to the literature of analytic metaphysics, "individuation" is understood in a metaphysical and an epistemic sense. In the epistemic sense, someone individuating an object "is to 'single out' that object as a distinct object of perception, thought, or linguistic reference." (Lowe 2005: 75) This epistemic sense presupposes the metaphysical sense, in which what 'individuates' an object "is whatever it is that makes it the single object that it is – whatever it is that makes it one object, distinct from others, and the very object that it is as opposed to any other thing." (Lowe 2005: 75) Bueno, Chen, and Fagan (2018) add a practical sense to the term, interpreting

"individuation" as a practical process through which an individual is produced. They characterize the relation between "individuation" and "individuals" as when "an individual emerges from a process of individuation in the metaphysical sense. Epistemic and practical individuation, then, are processes that aim to uncover stages of that metaphysical process." (Beuno, Chen, and Fagan 2018) The approach to the individuation of genes I adopt herein follows their characterization, especially by focusing on the process of epistemic and practical individuation. Reversely, the case I am investigating in this paper offer an illustration for the new sense of individuation.

Although philosophers have investigated concepts of the gene and its change by examining many cases in scientific practices, they have seldom considered the role that the transgenic technique developed in biotechnology may play in philosophical discussions. This paper explores experimental individuation of genes from the direction of that technique, considering the possibility that a gene is individuated as an individual in the relevant contexts.

This paper thus addresses two central questions: (Q1) In what sense, can we reasonably say that classical geneticists have individuated a gene? (Q2) Are there experiments that can individuate a gene as an individual? Some new questions such as the relationship between individuality and individuation will be derived from the answer to the two questions. This paper is thus structured in the following way.

In the second section, I review the literature about the concepts and references of genes. Section 3 argues that the answer to Q1 is that the geneticists individuate a gene as a type, because they used the chromosomal location technique. Section 4 argues that the answer to Q2 is the experiments that use the transgenic technique. The two answers indicate two different kinds of individuation: individuation of a type and individuation of an individual. This raises a new question about whether or not "individuation of a type" is a consistent phrase. In order to respond to this, section 5 discusses in what sense we individuate a type and compare between two kinds of individuation defined by two different kinds of experiments and techniques: the chromosomal locution of genes and the transgenic experiment. My argument thus involves the relationship between kind and individual in the context of experimentation. Given the new question, Section 6 argues that transgenic experiments can demonstrate a gene type by individuating its tokens, while gene mapping experiments in classical genetics only individuate gene types. Thus, a new gene conception, calling it "the transgenic conception of the gene," can be proposed. I further discuss the relationship among the classical gene concept, the molecular gene concept, and the transgenic conception. In the seventh section, I defend the thesis that practices of individuation in scientific investigations are prior to characteristics of individuality identified by traditionally metaphysical speculations.

2. Concepts and references of the gene

The rapid change of the gene concept has produced a large multitude of gene concepts that have bewildered scientists (Gerstein et. al. 2007; Pearson 2006; Stotz and Griffiths 2004). The confused situation has attracted many philosophers and scientists to provide clarifying analyses. Although scientists as well as philosophers have made endeavors to overcome the predicament, they are motivated differently. Scientists believe that they need a unified concept to help them conduct research and to communicate with each other, because, as developmental geneticist William Gelbert says, "it sometimes [is] very difficult to tell what someone means when they talk about genes because we don't share the same definition" (Pearson 2006: 401). Thus, most scientists seek to redefine the "gene" and tend to adopt a single preferred perspective on the gene concept, although they are well aware with the plurality of gene definitions (Wain et. al. 2002; Gerstein et. al. 2007).

Philosophers at different times have been interested in clarifying concepts of the gene and in investigating the patterns of associated conceptual change. In contrast to actual definitions used by working scientists, they often consider more abstract concepts of the gene that can guide several different definitions in the context of scientific research. Consequently, they conclude that it is almost impossible to find a unified concept of the gene, and hence they take different stances to respond to this situation (rf. Waters 2007). Some are gene skeptics (e.g., Kitcher 1992). Some take a dualistic position, such as Moss (2003), who distinguishes between Gene-P and Gene-D based on the fields in that gene concepts are applied. Some are pluralists, such as Griffiths and Stotz (2006, 2013), who differentiate between three senses of the gene: the instrumental gene, the nominal molecular gene, and the postgenomic molecular gene. Still others are both pluralists and pragmatists. Waters (2018) emphasizes that scientists do and should apply different gene concepts under various investigative contexts.

With some exceptions, few philosophers explore the reference problem of the term "gene". Although Fregean semantics holds that the sense/concept or intension of a name determines its reference or extension, the matter about how a sense determines the reference is not easily seen from the scientific context. The determination of a theoretical term's reference usually involves experimental procedures and techniques that should be investigated and analyzed. Weber (2005, ch.7) does impressive work by providing several reference-determining descriptions of the term "gene" in the history of genetics. Based on those descriptions and the analysis of *Drosophila* genetic practices, he suggests that the pattern of referential change for "gene" is a kind of

freely floating reference. He also argues that different gene concepts refer to *different* natural kinds, which are overlapping but not coextensive.[1] According to Weber, reference for "gene" is fixed in the following manner for classical and molecular genes.

> Reference of [classical] "gene" (2): Whatever (a) is located on a chromosome, (b) segregates according to Mendel's first law, (c) assorts independent of other genes according to Mendel's second law if these other genes are located on a different chromosome, (d) recombines by crossing-over, (e) complements alleles of other genes, and (f) undergoes mutations that cause phenotypic differences. (Weber 2005: 210)

> Reference of [molecular] "gene" (5): The class of DNA sequences that determine the linear sequence of amino acids in a protein. (Weber 2005: 212)

Both classical and molecular gene concepts do refer to natural objects, because, as Weber notes (2005: 210-211), some *tokens* satisfying the reference-determining descriptions are experimentally presented when using the concepts with the intention of referring to sets of entities in historical occasions. However, one should note that the experimented tokens in classical genetics seems to be only some organisms with specific phenotypes (say, fruit flies or other kinds of organisms) while the experimented tokens in molecular biology may be some DNA segments. This difference raises interesting problem: what tokens are individuated in different contexts of experiments?

Before moving to the next section, I want to clarify that the individuation problem of gene concept's tokens is not the issue of gene individuality as raised by Rosenberg (2006: 121-133).[2] He defends the gene individuality thesis in parallel to the species individuality thesis, but Reydon (2009) objects to his argument and defends the gene as a natural kind. This paper aims to discuss how a gene kind and its tokens are individuated rather than whether or not an allele such as *Hbf* (the human fetal hemoglobin gene) is an individual.

3. Chromosomal location of a gene

[1] Baetu (2011: 411) argues that "the referents of classical and molecular gene concepts are coextensive to a higher degree than admitted by Waters and Weber…" However, Baetu builds his argument in terms of Benzer's work on phage. In my view, he does not successfully refute Waters' and Weber's arguments, because the referential change occurred within the classical gene concept, as Weber cogently argues.

[2] Rosenberg uses "natural selection and the individuation of genes" as the title of the section in which he discusses the gene individuality thesis.

Weber's argument indicates that we may and should consider the reference of the classical gene concept independently of the molecular gene concept and others. Weber's reference-determining description of "gene" (2) indicates that the chromosomal location (or mapping) of genes plays a key role in determining referents. However, the question "what tokens are individuated and thus referred to?" does not be answered.

Classical geneticists in the early 20th century located and labeled some specific classical genes on some specific chromosomes. The earliest genetic map (see Figure 1) of *Drosophila melanogaster* (fruit fly) was depicted in 1915. Figure 1 shows that the gene (allele) pair of *Drosophila's* grey body and (mutant) yellow body is located at the first locus on the first chromosome. The second gene pair of red eyes and (mutant) white eyes is located below the grey body gene. The other genes are located below the first two in order. However, every gene is differently distant from the first gene and thus occupies *a single locus* without overlapping. Accordingly, are we able to say that the location of a gene individuates the gene? Before answering this question, it is necessary to discuss how classical geneticists locate a gene on a chromosome. In other words, what technique is used in the process of locating genes?

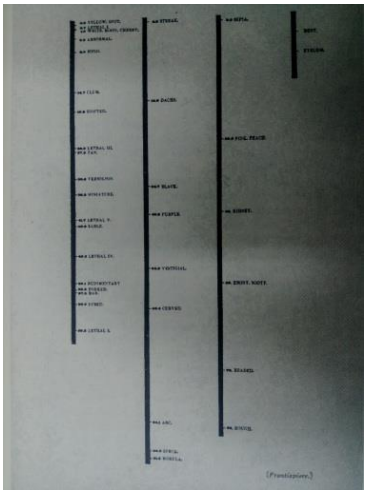


Fig. 1. Genetic map of *Drosophila* in 1915. Reproduced from Morgan, T. H. et. al. (1915).

Chromosomal location or mapping of genes is a well-known story (Darden 1991, Waters 2004, Weber 2005, 2006; Falk 2009). For the purpose of this paper, I introduce a very brief version. In the 1910s, Thomas Hunt Morgan's team developed a

technique to map the linear relations among factors (genes) in linkage groups, using Mendelian breeding data. Morgan and his team discovered that a pair of chromosomes may cross over with each other partially during the period of meiosis. Crossing over produces a specific ratio of the linked traits. Morgan believed that "the percentage of crossing over is an expression of the 'distance' of the factors from each other." (Morgan et.al. 1915: 61) Sturtevant then used percentages of linked characters that exhibited crossing over to calculate the relative positions of the factors to each other. This is the kernel technique for constructing genetic maps. By using genetic maps, Morgan's team determined the loci of many genes on the four chromosomes of *Drosophila*. Given the genetic maps, the classical geneticists assume that no other genes are located at the same position of a chromosome.[3] As a consequence, the single location of a gene actually indicates the individuality of genes.

Genetic maps by nature are diagrammatic models for the actual loci of genes in chromosomes. They are inferences from the statistical data of breeding experiments. Models represent the general. When we say that the location of a gene in a genetic map represents the locus of a classical gene on a chromosome, we really mean that it represents the locus of a type of classical gene on an identical type of chromosome in a cell within a kind of organism. Of course, this implies that a token of a type of classical gene on a token of a type of chromosome can be cognitively identified and discerned, because we can distinguish it from the tokens of the other genes. As a result, we can also count genes within cells. The located genes thus satisfy the two traditional characteristics of individuality: distinguishability and countability.[4]

If all chromosomes were stick-shaped substances of uniform material without complicated structure, then the chromosomal location of classical genes would be able to genuinely individuate them. According to molecular biology, however, chromosomes are a long chain of double helix DNA molecules that curl themselves up in twisted shapes. In such a case, we cannot delineate a located classical gene or depict its contour or boundary, because the chromosomal locus at which the gene is located includes a twisted part of the long DNA molecule. Even by invoking the knowledge from molecular biology, one would still be puzzled by the problem of defining the molecular gene.

4. Individuating molecular genes as individuals

Ever since the era of molecular biology, the continuously accumulating knowledge of genetics has not solved the individuation problem of genes. Instead, it

[3] Of course, a full story is more complicated. For the simplifying purpose, I skip the relevant discussion about gene mutation.

[4] The implications of using these criteria will be discussed in the sixth section.

has brought more troubles about the definition of the gene concept. Is a gene "a sequence of DNA for encoding and producing a polypeptide"? Should we include the start and stop codons (i. e., the regulation problem)? Should we count those introns deleted during the process of transcription into the investigated gene (i.e., the splicing problem)? The difficulty in defining the molecular gene concept directly contributes to the impediment of individuating a gene.

Many gene sequencing projects have been conducted during the genomic era. Scientists do not identify a DNA sequence as a gene and discern the gene from others by using gene sequencing *per se*, because it offers only syntactical orders of genetic codes. Gene annotation, which is used to infer what those annotated sequences do, has been developed to offer *senses* or *intensions* for them. However, the impediment of discerning genes remains, because the definition of the gene is still vague and confusing (rf. Baetu 2012; Gerstein et. al. 2007; Griffiths and Stotz 2013, ch. 4). In fact, gene annotation is based on several assumptions, by which scientists infer that a few sequences may be genes that contribute to phenotypes or functions. Those assumptions need to be confirmed by experimental investigations. Many techniques such as directed deletion, point mutation making, gene silencing, and transgenesis in reverse genetics have been developed to determine what a gene is and what it does (Gilchrist and Haughn 2010).

I argue that the transgenic technique is a very definite and powerful way to individuate a gene. It can even individuate molecular genes as individuals without a clear boundary of a gene or a clear definition of the gene, although the technique is limited.[5] How does the transgenic technique do this? What conditions of individuality allow the technique to individuate a gene as an individual?

Chen (2016) proposes a conception of experimental individuality with three attendant criteria (separability, manipulability, and maintainability of structural unity) and argues that the first experiment of bacteria transformation individuated an antibiotic resistance gene by satisfying the three criteria.[6] Below I reiterate this story in brief.

Stanley Cohen and Herbert Boyer combined DNA of *Escherichia coli* (*E. coli*) in 1973 and 1974 by transferring two different DNA segments encoding proteins for ampicillin and tetracycline resistance into *E. coli*, thereby realizing the transformation of this bacterium (Cohen et. al. 1973; Chang and Cohen 1974). Both DNA segments are called an "antibiotic resistance gene." Cohen and Boyer used small circular

[5] The technique cannot be applied in many occasions because of technological difficulties. It should not be applied to humankind due to ethics consideration. In addition, many gene-modification organisms produced by using the technique may involve ethical issues.

[6] Chen (2016) uses the creation of Bose-Einstein condensates in physical experiments as the other example. Chen's intent is to argue that biological entities and physical entities in laboratories share the same criteria of experimental individuality.

plasmids (extrachromosomal pieces of DNA) as vectors to transfer a foreign DNA segment into a bacterial cell. The plasmids were made by cutting out a (supposed) antibiotic resistance gene from other bacteria with the restriction enzyme *EcoRI*, linking the segment into a plasmid by using another enzyme, DNA ligase. The scientists then transferred the plasmid into an *E. coli* cell without the ability to resist antibiotics. The result, a modified *E. coli* cell, was able to resist antibiotics and contained the antibiotic resistance gene. In that experiment, the antibiotic gene was separated from its original bacteria and then was manipulated (i.e., linked and transferred). Its structural unity was not broken down, hence allowing it to be expressed in the other kind of bacteria. Scientists thus identify it as a gene, an individual biological entity, because the separated, manipulated, and maintained antibiotic gene was naturally separable, manipulable, and maintainable. The photos in Figure 2 show that scientists worked with a single DNA segment, as indicated by (b) in [A] and [B].

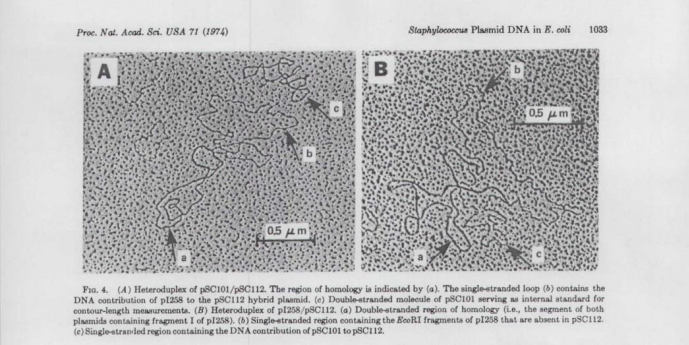


Fig. 2. Two pictures of plasmids in bacterial transformation. Reproduced from Chang and Cohen (1974).

I next interpret the performance of the technique used in transgenic experiments as the general process of individuating transgenes. The process has five stages.

(1) Use restriction enzymes to cleave specific segments from recognition sites of long DNA chains. A specific restriction enzyme can cut away a specific DNA segment at a specific site.

(2) Link the cleaved segment of DNA to a plasmid vector by using DNA ligase. The vector is a circular DNA that may come from a wild type of virus.

(3) Incorporate the DNA segment in the vector into the genome of another organism by injecting the plasmid vector to a cell of the target organism. Of course,

they may fail when the intended feature is not expressed.

(4) Make copies of DNA segments by cloning the cell containing the transferred segment of DNA. The aim of DNA cloning is to copy a segment of interest (or a gene) from an organism and produce many copies.

(5) Observe the expression of the novel feature that the target organism does not typically have. If a DNA segment cut from an original organism is successfully pasted into a cell of a target organism and the target organism expresses the intended feature that the original organism has, then one concludes that the segment is a gene.

The first stage corresponds to the separation condition, the second, the third, and the fourth stages to the manipulation condition, and the fifth stage to the maintenance condition. Accordingly, one can easily see that those cut, linked, transferred, pasted, and copied genes are particulars – individuals, because they satisfy the three criteria of experimental individuality that indicates their *singularity* and *particularity*. In other words, a single segment of DNA maintains its structural unity when being separated and manipulated. This is so, because cutting a gene from an original organism is in fact separating it from its environment and because transferring, pasting and copying a gene is manipulating it. If the gene does express the intended feature in a target organism, then this condition indicates that the unity of its chemical and informational structure has been maintained.

5. Two kinds of individuation of genes

The previous discussion indicates that two different objects have been individuated in different experimental and theoretical contexts. In the context of classical genetics, scientists used breeding experiments and theoretical inferences to locate a gene at some locus on a chromosome. They would individuate genes as types if they assume that no other genes could coexist at the same locus. If one interprets the meaning of "individuation" as "only individuals can be individuated," then the phrase "individuating genes as types" sounds unreasonable. Is it better to say "unitization of genes" rather than "individuation of genes"?

It is quite right to say classical geneticists *unitize* genes as types. In a sense, however, we may reasonably say that we individuate a gene as a type, because the type has tokens or members that are distinguishable and countable individuals. Classical geneticists suppose that all types of genes have corpuscular members, i.e., substantive individuals. In such a sense, talking of "individuating genes as types" is reasonable. If no distinguishable and countable members or samples of a kind can be identified, then the kind cannot be individuated. In other words, we cannot individuate

such a kind as water or air that is expressed by "mass" nouns at the macroscopic level, although we can individuate a sample of water by using a container or individuate a water molecule by specific technique at the molecular level. For the cases of experiments using the transgenic technique, molecular biologists physically individuate *singular and particular* gene tokens. Thus, we claim that scientists experimentally individuate genes as individuals in such a context.

In consequence, two different sets of criteria for individuality are presupposed. Experiments using the location technique have individuated a type whose tokens or members are countable individuals rather than matter referred to by mass nouns. In such experimental contexts, we emphasize distinguishability and countability as the indexing features of individuals. Experiments using the transgenic technique individuate singular and particular individuals – gene tokens. For these experimental contexts, we emphasize singularity and particularity of individuals in contrast to universality of types or kinds. We assure the particularity and singularity of the individuals through the realization of experimental individuality, namely, the joint realization of separability, manipulability, and maintainability of structural unity. At this point, more philosophical implications will be discussed in next section.

The two individuated targets indicate two different referential levels of the term "gene" in the literature. As we have seen, when many philosophers and scientists ask "what is a gene," they really refer to a type of gene in conjunction with discussing the gene concept or the definition of "gene." Similarly, in some contexts of scientific investigation, scientists use "a gene" to refer to a type of gene as the phrase "chromosomal location of a gene". In the context of transgenic experiments, however, "a gene" is used to refer to a genuine individual – a single and particular gene token, because scientists have worked with particular objects that maintain their structural unity when being separated and manipulated in the process of experimenting.

The two referential levels indicate two different kinds or levels of experimental individuation, which are realized by two different techniques: the chromosomal location technique and the transgenic technique. Although the two techniques aim to the same target (i.e., genes or types of genes), they physically experiment and manipulate different objects. Experiments using the chromosomal location technique indirectly identify loci of genes by manipulating organisms that contain chromosomes with genes in breeding, while experiments using the transgenic technique directly manipulate DNA segments. Therefore, classical geneticists can only cognitively discern gene types by identifying their loci without practically interacting with gene tokens; they really practically interact with organismal individuals that contain different types of genes. Reversely, molecular biologists can practically interact with gene tokens and then cognitively infer out the existence of a gene type.

6. Gene concepts and individuation

One may still wonder: Can the location technique individuate a singular and particular gene in the sense of individuating entities as individuals? The answer is obviously negative, because that technique cannot separate and manipulate a gene token and maintain its structural unity. On the contrary, one may ask: Can the transgenic technique individuate a type of gene? Here the answer is less clear. In the sense that scientists suppose that a token of a gene has been physically individuated in transgenic experiments, we are allowed to say that the technique also individuates a type of gene. However, scientists are not fully sure that the transgenic technique on a posited gene can be always successfully applied to another individual of the same organism. In fact, the probability of failure is quite high. Unless the experimental individuation of particular tokens can be performed repeatedly and stably, then one can say that the gene tokens indicate a general type of gene and that the type has been identified. However, the object individuated by the technique is not a type of gene, because the technique always requires manipulating particular segments of DNA -- gene tokens. If a kind of transgenic experiment with a specific transgene has been stably repeated, then a type of gene has been discovered by experimentally individuating its tokens in performing such an experiment.

Since transgenic experiments may be successfully and stably performed by using different transgenes, one can extract a special conception of the gene that is characterized by the transgenic technique. I call this "the transgenic conception of the gene," in which *a gene is a transferrable DNA sequence which is able to express a phenotype/function on another kind of organisms*. Of course, this does not imply that those technically untransferrable DNA sequences are not genes, given the fact that the number of transgenes is relatively few to the number of genes located at chromosomes. This is so because scientists do not always find the precise site of a gene (type) and available restriction enzymes to cut the DNA segment of the gene. Thus, the extension of the transgenic conception of the gene is not equivalent to that of the classical gene concept. Due to the limited number of transgenes, the transgenic conception is not yet co-extensional with the molecular gene concept. To be precise, the extension of the former is included within the extension of the latter, because all transgenes are molecular genes but not all molecular genes can be transplanted. In addition, the intension of the transgenic conception is implied in the intension of the molecular gene concept, because the technique was developed from molecular biology. As a consequence, the transgenic conception can be viewed as a *sub-conception* of the molecular gene concept. Nevertheless, we have a conception

derived from scientific practices.

7. The priority of individuation to individuality

Bueno, Chen, and Fagan (2018) promote an approach by which investigating processes of individuation in scientific practices is prior to metaphysical speculation on criteria of individuality. This paper obviously follows the approach. However, this does not mean that we do not need any criterion of individuality in identifying any individual in scientific practices. Rather, criteria of individuality are implied in or extracted from procedures of scientific practices, as the three conditions of experimental individuality are extracted from experimental practices (Chen 2016). Criteria of individuality based on scientific practices may or may not conflict with criteria from metaphysical theories. Considering the relationship between practical criteria and speculative criteria will help us understand practical individuation more deeply.

The metaphysical tradition has identified at least six characteristics or indexing features of individuality in general: particularity, distinguishability, countability, delineability, unity, and persistence (Pradeu 2012: 228-229; Chen 2016: 351).[7] Recently, some philosophers argues that all biological entities are processes (Dupré 2018, Nicholson and Dupré 2018, Pemberton 2018), so I would like to add processuality to the list. Indeed, I believe that all biological individuals pass through a life, i.e., a process (see also Chen 2018), therefore, processuality is a central characteristic of biological individuality. Those characteristics, originally come from metaphysical speculation, can singly, jointly, or collectively serve as epistemic criteria of individuality.

In the context of scientific practices, they are the outcomes from rather than preconditions for the realization of individuation. For example, individuating genes as individuals in the context of transgenic experiments indicates that the separated, manipulated, and maintained genes are particular and singular tokens. As the experimental individuation of gene tokens is realized, those tokens are also distinguishable, countable, unitary, persistent, and passing through a process, because particular and concrete individuals are being separated, manipulated, and maintained. The practices of separation and manipulation indicate epistemic particularity,

[7] Characteristics of individuality can serve as criteria of individuality and thus be involved in a theory of individuation. Bueno, Chen, and Fagan (2018) identify six theories of individuation in traditionally analytic metaphysics. A theory of individuation in the metaphysical sense involves not only "a theoretic construction of the nature of individuality and its attendant criteria," but also other metaphysical concepts such as "property, trope, universal, particular, substance, substratum, time, space, sort or kind." (p. 3) For my purpose, I will discuss only characteristics of individuality rather than any theory of individuation.

distinguishability, and countability. The practice of maintenance of structural unity indicates the unity, persistence, and processuality of the maintained gene token. However, all of the three practices would not indicate the delineation of a gene token, because the spatial boundary of the manipulated gene does not and cannot be delineated. Of course, this point does not mean that delineation is not a characteristic of individuality, but rather that it is not applicable to this case.

Individuating genes as types in classical genetics indicates that the individuated types of genes contain distinguishable and countable tokens, because the individuation is the location of a gene at a chromosome in a diagrammatic model. Supposing that the loci of different genes do not overlap, then the special locus of a gene is thus distinguishable from the locus of another gene. As a consequence, a gene token at a chromosome in a cell of a kind of organism is thus distinguishable from another token of the identical type of gene. All gene types located at chromosomes are countable. Supposing that every organism contains a token of a specific type of gene, then tokens of that gene type are countable. However, chromosomal location of genes does not indicate particular and singular gene tokens, because the individuated objects are only types of genes. As I have argued, the kind of individuation practice did not touch down the manipulation of individuals and remained in the cognitive level which focuses on gene types in general.

Although the concept of individuation can be reasonably applied to a kind whose members are individuals, all characteristics of individuality are not applicable. One cannot apply particularity, delineation, unity, and processuality to gene types, because a gene type is, in principle, universal, occupying multiple spaces, not cohesive, replicable, and non-processual. However, distinguishability and countability can be adequately applied to gene types, because one can distinguish one gene type from another gene type and count gene types when the chromosomal location is realized. In this case, thus, both distinguishability and countability cannot sufficiently demonstrate that the individuated objects are individuals. On the other hand, in the case of transgenic experiments, we can derive particularity, unity, and processuality from the three conditions of experimental individuation (separation, manipulation, and maintenance of structural unity). As a consequence, characteristics of individuality are derived from individuation; they are outcomes of practical individuation.

8. Conclusion

In this paper, I argue that there are at least two kinds of experimental individuation of genes. Scientists individuate genes as types in classical genetics and

individuate genes as tokens in transgenic experiments. Individuating a gene as a type or individuating a gene as an individual depends on the technique used in experimentation. I argue that characteristics of individuality identified in traditional metaphysics are not presupposed by individuation. Rather, they are outcomes or products derived from practical individuation in scientific experiments. I further argue that different kinds of experimental individuation presuppose different concepts of the gene: the classical gene concept and the transgenic conception of the gene. I argue that the transgenic conception can be viewed as a sub-conception of the molecular gene concept. An outstanding problem remains. Whether we can unify different concepts of the gene by integrating different experimental techniques, such as the chromosomal location technique, the technique of genetic sequencing, the techniques in reverse genetics, and the transgenic technique. Future analyses can approach this and other related questions in light of our new understanding of how classical geneticists individuated genes and the role experimental techniques play in identifying a gene as an individual.

Acknowledgment: I thank Alan Love, Ken Water, and Marcel Weber for their very valuable comments and suggestions. This paper has been revised according to their comments.

References

Baetu, Tudor M., 2011. “The referential convergence of gene concepts based on classical and molecular analysis,” *International Studies in the Philosophy of Science*, 24 (4): 411-427.

Baetu, Tudor M., 2012. “Genes after the human genome project.” *Studies in History and Philosophy of Biological and Biomedical Science*, 43: 191-201.

Beurton, P., R. Falk, and H.- J. Rheinberger, 2000. *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*. Cambridge, UK: Cambridge University Press.

Beuno, Otavio, Ruey-Lin Chen, and Melinda B. Fagan, 2018. “Individuation, process, and scientific practice.” In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 1-18. New York: Oxford University Press.

Carlson, E., 1991. “Defining the gene: an evolving concept.” *American Journal of Human Genetics*, 49: 475-487.

Chang, Annie C. Y. and Stanley N. Cohen, 1974. “Genome construction between bacterial species *in vrito*: Replication and expression of *Staphylococcus* plasmids

in *Escherichia coli*," *Proceedings of the National Academy of Science of USA*, 71(4): 1030-1034.

Chen, Ruey-Lin, 2016. "The experimental realization of individuality." In Alexandre Guay and Thomas Pradeu (eds.). *Individuals across the Sciences*, 348-370. New York: Oxford University Press.

Chen, Ruey-Lin, 2018. "Experimental Individuation: Creation and Presentation," In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, . New York: Oxford University Press.

Cohen, Stanley N. et. al., 1973. "Construction of biologically functional bacterial plasmids *in vitro*," *Proceedings of the National Academy of Science of USA*, 70(11): 3240-3244.

Darden, Lindley, 1991. *Theory Chang in Science: Strategies from Mendelian Genetics*. Oxford: Oxford University Press.

Dupré, John, 2018. "Processes, Organisms, Kinds, and Inevitability of Pluralism." In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 25-38. New York: Oxford University Press.

Falk, Raphael, 1986. "What is a gene?" *Studies in History and Philosophy of Science*, 17: 133-173.

Falk, Raphael, 2009. *Genetic Analysis: A History of Genetic Thinking*. Cambridge: Cambridge University Press.

Falk, Raphael, 2010. "What is a gene – revised" *Studies in History and Philosophy of Biological and Biomedical Science*, 41: 396-406.

Gerstein, Mark B. et. al. 2007. "What is a gene, post-ENCODE? History and updated definition." *Genome Research*, 17(6): 669-681.

Gilchrist, Erin and George Haughn, 2010. "Reverse genetics techniques: engineering loss and gain of gene function in plants," *Briefings in Functional Genomes*, 9(2): 103-110.

Griffths, Paul and Karola Stotz, 2006. "Genes in the postgenomic era," *Theoretical Medicine and Bioethics*, 27(6): 253-258.

Griffths, Paul and Karola Stotz, 2013. *Genetics and Philosophy: An Introduction*. Cambridge: Cambridge University Press.

Kitcher, P. S., 1982. "Genes." *British Journal for the Philosophy of Science*, 33: 337-359.

Kitcher, P. S., 1992. "Gene: current usages." In E. Keller and L Lloyd (eds.), *Keywords in Evolutionary Biology*. Cambridge, MA: Harvard University Press, pp. 128-131.

Lowe, E. Jonathan 2005. "Individuation," *The Oxford Handbook of Metaphysics*, ed. Michael J. Loux and Dean W. Zimmerman. Oxford: Oxford University Press.

Maienchein, J., 1992. "Gene: Historical perspectives." In E. Keller and E. Lloyd (eds.). *Keywords in evolutionary biology*. Cambridge, MA: Harvard University Press, pp. 181-187.

Morgan, Thomas Hunt, et.al., 1915. *The Mechanism of Mendelian Heredity*. New York: Henry Holt and Company.

Moss, Lenny, 2003. *What Genes Can't Do*. Cambridge, Mass.: The MIT Press.

Nicholson, Daniel J. and John Dupré, 2018. Everything flows: Towards Processual Philosophy of Biology.

Pearson, Helen, 2006. "What is a gene?" *Nature*, 441(25): 399-401.

Pemberton, John. 2018. "Individuating Processes," In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 39-62. New York: Oxford University Press.

Pradeu, Thomas, 2012. *The Limits of the Self: Immunology and Biological Identity*. Oxford: Oxford University Press.

Reydon, Thomas, 2009. "Gene Names as Proper Names of Individuals: An Assessment." *British Journal for the Philosophy of Science*, 60(2): 409-432.

Rosenberg, Alexander, 2006. *Darwinian Reductionism*. Chicago: The University of Chicago Press.

Snyder, Michael and Mark Gerstein, 2003. "Defining genes in the genomics era." *Science*, 300(5617): 258-260.

Stotz, Karola and Paul Griffths, 2004. "Genes: philosophical analyses put to the test." *History and Philosophy of the Life Sciences*, 26: 5-28.

Wain, H. M., et. al. 2002. "Guidelines for human genome nomenclature," *Genomics*, 79: 464-470.

Waters, Kenneth C., 1994. "Genes made molecular," *Philosophy of Science*, 61: 163–185.

Waters, Kenneth C., 2004. "What was classical genetics?" *Studies in History and Philosophy of Science*, 35 (4): 783-809.

Waters, Kenneth C., 2007. "Molecular genetics," *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/molecular-genetics/

Waters, Kenneth C., 2018. "Don't Ask 'What is an individual?'" In Otavio Beuno, Ruey-Lin Chen, and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 91-113. New York: Oxford University Press. (In press)

Weber, Marcel, 2005. *Philosophy of Experimental Biology*. Cambridge, UK: Cambridge University Press.

Weber, Marcel, 2006. "Representing genes: Classical mapping techniques and the growth of genetic knowledge," *Studies in History and Philosophy of Biological and Biomedical Science*, 29: 295-315.

The Verdict's Out:

Against the Internal View of the Gauge/Gravity Duality

4993 words

Abstract

The gauge/gravity duality and its relation to the possible emergence (in some sense) of gravity from quantum physics has been much discussed. Recently, however, Sebastian De Haro (2017) has argued that the very notion of a duality precludes emergence, given what he calls the internal view of dualities, on which the dual theories are physically equivalent. However, I argue that De Haro's argument for the internal view is not convincing, and we do not have good reasons to adopt it. In turn, I propose we adopt the external view, on which dual theories are not physically equivalent, instead.

1 Introduction

The gauge/gravity duality has generated much discussion about whether space-time geometry or gravity emerges (in some sense) from quantum physics.[1] Recently, however, De Haro [2017] has argued that the very notion of a duality *precludes* the possibility of emergence given what he calls the *internal view* of dualities, on which dual theories are physically equivalent. In turn, this claim impinges upon the broader debate about whether we can make claims about emergence given a duality. After all, since the internal view of dualities is supposed to *rule out* emergence, any such debate is rendered moot once we adopt the internal view. My goal here, though, is to argue that De Haro's argument for the internal view is not convincing. Instead, I propose we adopt the *external view* of dualities, on which dual theories are *not* physically equivalent.

First, I introduce Fraser's [2017] three-pronged distinction of predictive, formal and physical equivalences, characterizing dualities in terms of this distinction (§2.1). I then make things more concrete by briefly considering the gauge/gravity duality via the Ryu-Takayanagi conjecture from the **AdS/CFT** (anti-de Sitter space/conformal field theory) correspondence (§2.2).

Next, I introduce De Haro's interpretive fork between the internal and external views of dualities (§3). I illustrate how the internal view is supposed to preclude emergence, but criticize De Haro's argument for the internal view – that it is meaningless to hold the external view given 'some form of' structural realism and how the two theories are

[1] One prominent physicist who is a proponent of emergent space-time is Seiberg 2007, while philosophers like Rickles 2011/2017, Teh 2013, and Crowther 2014 have all tackled the topic.

'totalizing' in some way – by showing how it does not work without further assumptions (§4). In turn, given the interpretive fork, I propose we adopt the external view instead. In concluding remarks, I briefly discuss this result in relation to the broader debate about emergence within the gauge/gravity duality.

2 Gauge/Gravity through AdS/CFT

2.1 Duality

Fraser [2017] takes two theories related by a duality to have two features: (i) they agree on the transition amplitudes and mass spectra, and (ii) there is a 'translation manual' that allows us to transform a description given by one theory to a description given by another theory. We may explicate (i) and (ii) by first considering distinct sorts of 'equivalence' proposed by Fraser [2017, 35]:

- *Predictive equivalence*: "there is a map from T_1 to T_2 that preserves the values of all expectation values deemed to have empirical significance by T_1 and that preserves the mass spectra, and vice versa."
- *Formal equivalence*: "there is a translation manual from T_1 to T_2 which maps all quantum states and quantum observables deemed to have physical significance by T_1 into quantities in T_2 and respects predictive equivalence, and vice versa."
- *Physical equivalence*: "there is a map from T_1 to T_2 that maps each physically significant quantity in T_1 to a quantity in T_2 with the same physical interpretation and respects both formal and predictive equivalence, and vice versa."

Given our characterization of a duality as (i) and (ii), we may quite naturally say that two theories are dual to one another when they are *predictively* and *formally* equivalent. Furthermore, supposing that this three-pronged distinction exhausts the possible equivalences relevant to physics, we might also say that two theories satisfying (i)-(iii) are also *fully*, or *theoretically*, equivalent.

Here it would be germane to differentiate two distinct sorts of structures in a duality. Given predictive and formal equivalence, the isomorphism holding between physical and empirical quantities of the dual theories suggests a structure, which may be called the *empirical core* of the duality. However, as Teh [2013, 301] also notes, despite the empirical core, "duality is precisely an equivalence between two theories that describe (in general) different physical structures, i.e. theories with non-isomorphic models." In other words, while there is an empirical core, by which physical and empirical quantities are mapped onto one another, these quantities are generally related to other quantities in a quite different manner on each side, viz. there is 'excess structure' exogenous to the empirical core. Without further argument, we are not entitled to 'discard' this 'excess structure', which also means that predictive and formal equivalence (characterizing the empirical core) does not automatically entail physical, and hence full, equivalence.

Given Fraser's framework, I will briefly introduce the gauge/gravity duality more concrete by briefly examining the example of **AdS/CFT** correspondence.

2.2 The AdS/CFT Correspondence

The *gauge/gravity duality*, or *holographic principle*, postulates a duality between a suitably chosen N-dimensional gauge quantum field theory (QFT) that does not describe

gravity, and a quantum theory of gravity in (*N+1*)-dimensional space-time (the 'bulk') with an *N*-dimensional 'boundary', on which the gauge theory is defined. Hence the slogan: gauge on the boundary, gravity in the bulk.

The **AdS/CFT** correspondence is a specific case of the gauge/gravity duality. On the one hand, '**AdS**' stands for anti-de Sitter space-time - a maximally symmetric solution to the Einstein equations with a constant negative curvature and a negative cosmological constant. More accurately, though, the '**AdS**' in **AdS/CFT** correspondence should be taken to refer to a string theory of quantum gravity defined *on* a 5-dimensional **AdS**. '**CFT**', on the other hand, refers to a quantum field theory with scale (or conformal) invariance defined on the 4-dimensional boundary of the **AdS**. The **AdS**-side theory is defined in the 'bulk', and the **CFT**-side theory is defined on the 'boundary' of the **AdS** space-time.

The **AdS/CFT** correspondence, then, refers to a postulated duality between the two theories, satisfying (i) and (ii) from §2.1. (i) is satisfied given the postulate that bulk fields propagating in the bulk are coupled to operators in the boundary **CFT**. Hence, the **AdS** theory of gravity will predict exactly the 'same physics', viz. transition amplitudes, expectation values and so on, as the **CFT** theory without gravity.

Beyond empirical, i.e. measurable, quantities, physically significant quantities of **AdS/CFT** must also relate to one another since it is a duality. In other words, (ii) is supposed to hold simply as a core postulate. This is not to say that (ii) is completely unfounded: in particular, we have evidence suggesting that at least *some* physical quantities of dual theories are related to one another in surprising ways, which in turn supports the claim that (ii) holds. Here I will focus on one such relation, the Ryu-Takayanagi conjecture.

The Ryu-Takayanagi conjecture postulates that the entanglement entropy of two regions on the boundary is related to the surface area within the bulk:[2]

$$(\mathbf{RT}):\ S_A = \frac{Area(\widetilde{A})}{4G_N}$$

RT tells us that the entanglement entropy of a region on the boundary of the **AdS**, S_A, viz. the von Neumann entropy[3] in the **CFT**, is directly proportionate (by 4 times the Newtonian gravitational constant) to the area of the boundary surface $\widetilde{A}$ bisecting the bulk, dividing the two entangled regions on the boundary. Below, *Fig. 1.* shows a simplified diagram for visualizing **RT**.

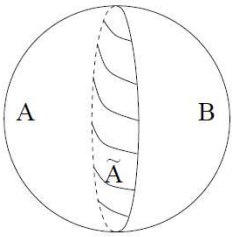


Fig. 1. The area $\widetilde{A}$ bisects the bulk space-time into two, and on the boundaries of the two parts we define the regions A and B. The Ryu-Takayanagi formula tells us that given a change in S_A we get a change in the size of $\widetilde{A}$ by the proportion of $\frac{1}{4G_N}$. [Figure taken from Van Raamsdonk 2010]

RT paints an interesting picture for emergence of space-time geometry from quantum theory: the area of a space-time itself is closely related to quantum entanglement entropy in a surprising way. An increase in the entanglement entropy between two

[2]See Ryu & Takayanagi 2006 for technical details.

[3]The von Neumann entropy is given by $S_A = -Tr(\rho_A log \rho_A)$. The reduced density matrix describing the region A, ρ_A, is obtained from tracing over the B-components of the combined density matrix of A and the entangled region B, ρ_{AB}: $\rho_A = Tr_B(\rho_{AB})$.

regions of a field described by **CFT** leads to a proportionately increasing boundary area of the bulk, and hence a geometric (or gravitational) phenomenon is described in terms of a quantum phenomenon.[4]

Given relations like **RT**, we can also see more clearly how **AdS/CFT** is supposed to satisfy (ii): physically significant quantities, such as 'area' of space-time in the bulk and 'entanglement entropy' between two regions on the boundary, are mapped to one another via suitable equations. Hence, **AdS/CFT** is a special case of the gauge/gravity duality: a theory of quantum gravity on a $(N+1)$-dimensional **AdS** space-time is dual to a **CFT** defined on its N-dimensional boundary.

With the gauge/gravity duality made concrete, let us turn to the interpretive task.

3 The Internal View

Dieks et al. [2015] and De Haro [2017] proposes an interpretive fork for dualities: we can either adopt an internal or external view. De Haro describes the *internal view* as such:

> if the meaning of the symbols is not fixed beforehand, then the two theories, related by the duality, can describe the same physical quantities. [...] we have two formulations of one theory, not two theories. [De Haro 2017, 116]

On the contrary, the *external view* holds that:

> the interpretative apparatus for the entire theory is fixed on each side. [...] On this interpretation there is only a formal/theoretical, but no empirical, equivalence between the two theories, as they clearly use different physical

[4]See Van Raamsdonk 2010 for an excellent summary of this picture.

> quantities; only one of them can adequately describe the relevant empirical observations.

Is De Haro's characterization of the external view adequate? The fact that there is no 'empirical' equivalence (what Fraser calls physical equivalence) between two theories does not entail that at most one of them can adequately describe the relevant empirical observations, where one description is 'correct' and the other 'wrong', nor does it entail mutually exclusive physics where only one theory can be correct at any one time. To assume so seems to rule out, by fiat, the possibility of emergence, since emergence relies on *both* theories being in a way adequately descriptive of the world (except one is more 'fundamental' than the other). Hence, taking in account Fraser's framework, I re-characterize the external view as such: it is simply the claim that the two dual theories are *physically non-equivalent* i.e. have distinct physical interpretations, despite formal and predictive equivalence.

Given the interpretive fork, if we are led to forsake the internal view, then we are motivated to accept the external view instead. As such, my strategy here is to show that we should forsake the internal view, and in turn accept the external view instead.

To better understand what the internal view is claiming, I break it down into three constituent claims.

The first claim is that of *theoretical equivalence*: under the gauge/gravity duality, the two theories (e.g. **AdS** and **CFT**) are taken to be simply different formulations of a single theory, describing the same physical quantities despite their obvious differences. As Dieks et al. puts it, 'the two theories collapse into one' [2015, 209-210]. In light of Fraser's framework described in §2.3, this claim means that the gauge/gravity duality, on

the internal view, involves the conjunction of predictive, formal and physical equivalences. In other words, beyond a one-to-one mapping (a 'translation manual') of relevant physical quantities and the sharing of all transition amplitudes, mass spectra and other observable predictions, the internal view claims that the two theories also have the *same physical interpretation*. However, as Fraser [2017, 35] notes, "predictive equivalence does not entail formal equivalence, and formal equivalence does not entail physical equivalence." Formal and predictive equivalence cannot entail physical equivalence on their own.

The internal view's claim of theoretical equivalence, then, must require an additional claim of *physical equivalence*, in addition to formal and predictive equivalence: the dual theories are taken to be physically equivalent, and hence have the same physical interpretation. As per §2.1, this would indeed entail theoretical equivalence.

Physical equivalence is in turn justified by a third claim, that the two theories in a duality should be left *uninterpreted*. As De Haro claims above, assume 'the meaning of the symbols is not fixed beforehand'. Then, given formal and predictive equivalence, we have an isomorphism between the dual theories' (now-uninterpreted) 'physical quantities' and numerical predictions, viz. an uninterpreted empirical core. Ignoring the 'excess structure' exogenous to the empirical core, we can then take the empirical core to be representing a single uninterpreted theory, where the now-uninterpreted 'quantities' of each dual theory now refer to the 'places' or 'nodes' of the empirical core's structure. As Dieks et al. (2015) puts it,

> A in one theory will denote exactly the same physical quantity as B [...] if these quantities occupy structurally identical nodes in their respective webs

> of observables and assume the same (expectation) values. [Dieks et al. 2015, 209]

Now, given this situation, it might seem plausible to claim that the dual theories are really physically equivalent. Consider **RT**. On the internal view, we are led to say that 'area' really has the same meaning as 'entanglement entropy'. After all, in the theoretical structure that is supposed to matter on the internal view, viz. the empirical core, the two terms are related structurally in the same way to other terms elsewhere (sans a proportional constant). Given that the two theories is also stripped of all prior physical meaning, this structural identity suggests that the 'area' and 'entanglement entropy' are really describing the same quantities, despite their obvious non-isomorphism more generally (e.g. different equations in computing these quantities in their respective theories, the terms involved in calculating them, and so on). In other words, it seems that we are allowed to proclaim physical equivalence on this view.

If we do accept this third claim, we get physical and hence theoretical equivalence, and so the internal view does preclude the possibility of emergence: Theoretical equivalence effectively rules out any account of emergence. If the two dual theories are really just different formulations of one theory, then there is nothing for this new, unified, theory to emerge *from*: nothing can emerge from itself in any interesting way. Subsequently, a duality is supposed to *preclude* emergence on the internal view.

Agreed: physical equivalence entails theoretical equivalence, and theoretical equivalence rules out any sort of emergence. However, are we forced to adopt physical equivalence given the internal view? De Haro himself seems unclear on this point. Note the use of "can" in his characterization of the internal view above: "the two theories,

related by the duality, *can* describe the same physical quantities" [2017, 116, emphasis mine]. Are we supposed to believe that physical equivalence *can* hold, or that it *must* hold, on the internal view? In other words, since physical equivalence hangs on the third claim of leaving terms of the dual theories uninterpreted, *must* we adopt the third claim, or is it merely *possible*?

De Haro seems to suggest that theoretical, and hence physical, equivalence *must* hold, since he assumes the two dual theories to be 'two formulations of *one theory*' [emphasis mine]. However, later on, he suggests that physical equivalence merely *can* hold, when he considers an example of leaving dual theories uninterpreted beyond structural relations:

> For what might intuitively be interpreted as a 'length, a reinterpretation in terms of 'renormalisation group scale is now *available*.[5] [De Haro 2017, 116, emphasis mine]

The *availability* of an interpretative stance – in our case of **RT**, of interpreting bulk boundary surface area to be the same physical quantity as entanglement entropy – surely does not entail the *necessity* of the stance. Hence, there are two readings of the internal view: on the weak reading, we take the modal talk – e.g. a reinterpretation being 'available' or how we 'can' describe the same physical quantities – seriously, and on the strong reading we ignore the modal talk completely.

On the one hand, the claim that the internal view precludes emergence is not true on the weaker view. On this view, *if* we assume that the terms on both sides of the duality are uninterpreted, then there is no emergence; *but* this is not forced on us. In turn, this

[5]For context, though unmentioned in this paper, length and renormalisation group scale are also dual quantities in AdS/CFT.

makes the preclusion of emergence merely possible. However, this reading of the internal view does not rule out emergence as De Haro claims. I will thus assume that De Haro intends for us to take the strong reading of the internal view, which does claim that the terms of the both sides *are* uninterpreted.

However, we have not yet seen a compelling reason for accepting the claim that we *have to* see the terms of the dual theories as uninterpreted, and subsequently that physical equivalence *must* hold. *A fortiori* we are not obliged to accept the internal view.

Indeed, something is odd about the argument structure I mapped out: To establish the second claim of physical equivalence, we must establish the third claim, that we must discard anything beyond the empirical core and to leave the terms uninterpreted. However, to justify leaving the terms uninterpreted requires a convincing argument for assuming physical equivalence between the two theories to begin with! Otherwise, we have no reason to simply discard the 'excess' structure and leave the dual theories' terms uninterpreted.

Hence, further arguments are required to establish the third claim. Furthermore, if we discover that this argument is wanting, we shall then have reasons to reject the internal view.

4 De Haro's Argument

De Haro does provide an argument, which runs on the idea that two plausible commitments entails the internal view: the commitment that the dual theories are theories of the whole world in some suitably totalizing manner, and the commitment to "some form of structural realism" [2017, 116].

Let us begin by examining the two commitments. The first commitment implies that dual theories are theories of the whole world, in the sense that they are "both candidate descriptions of the same world" [Dieks et al. 2015, 14]. However, *prima facie* this is not true, since on one hand we have a theory of gravity/space-time geometry, while on the other we have a theory without (not to mention different dimensionalities). How can two theories, one describing something the other does not, both be about the same world? We can try to make this assumption intelligible by taking into account the translation manual between the two theories. Given the translation manual, we can claim that the **CFT** theory without gravity does describe gravity in a way. Consider **RT**: while the entanglement entropy described within **CFT** does not appear to describe space-time geometry *by itself*, the **CFT** plus the translation manual *and* **AdS** (in this case **RT**) *does* describe space-time geometry, albeit in a higher-dimensional space-time. When the entanglement described within the **CFT** changes, the boundary surface area in the **AdS**-side theory with gravity changes as well. Hence, by considering the translation manual given by the duality, the first commitment is made plausible.

The second commitment requires us to adopt some form of structural realism. Structural realism here can be understood loosely, since nothing turns on the particular account of structural realism we employ. Furthermore, De Haro himself does not specify precisely what he means by 'some form of' structural realism. As such, I will likewise adopt a loose notion of structural realism: I understand it to be the view that we should be (metaphysically or epistemically) committed only to the mathematical or formal structure of our theories, and this entails, among other things, that theoretical terms are to be defined in terms of their relations to other places or nodes in this formal structure.

Now, De Haro then claims that the two commitments entail the internal view:

> If [the two commitments] are met, it is impossible, in fact meaningless, to decide that one formulation of the theory is superior, since both theories are equally successful by all epistemic criteria one should apply. [De Haro 2017, 116]

Since he does not flesh out his argument in much detail, I attempt to reconstruct his argument in a plausible fashion: firstly, let us grant the two commitments. Do these commitments commit us to the conclusion that it is meaningless to differentiate between the two dual theories?

Dieks et al. [2015, 209] claims that given the first commitment, "it is no longer clear that there exists an 'external' point of view that independently fixes the meanings of terms in the two theories". However, I must admit I do not see why this is the case: as I explained above, the first commitment only makes sense *if* we understand both theories as having pre-determined meanings, and *then* relating them via the duality/translation manual. In other words, the first commitment is perfectly compatible with the external view.

For the remainder of this paper I focus on the second commitment instead. I think the second commitment *does* entail that differentiating the two theories is meaningless, *only if* we believe that one should be a structural realist (epistemically/metaphysically) only about the empirical core of the duality, discarding the 'excess structure' which made the two theories distinct structures to begin with. In other words, we want to say that this 'excess structure' was not physically significant to begin with: only the empirical core was relevant to physics. It seems that this is required to make sense of the claim that it is 'meaningless' to say that one formulation, e.g. the **CFT** side, is better than the

other, e.g. the **AdS** side. If structural realism commits us only to the empirical core of the dual theories, then accordingly there is really only one structure in question. Hence, it is meaningless to ask which structure is better (there is only one). If there is only one structure, then the internal view seems to hold: under a structural realist view, the terms of the dual theories are defined in terms of their places in the structure. Hence, within the empirical core's structure, the different terms of the dual theories really mean the same thing, and hence we get some version of the internal view.

Why should we, even as structural realists, commit ourselves only to the empirical core? The argument seems to me to be an epistemic one: we should believe that the structure relevant to the two theories given the duality must really be common to both theories because, as De Haro claims above, "both theories are equally successful" by all epistemic criteria we apply. If this is true then it seems we have no way of differentiating between the two theories, and the best explanation for this epistemic equivalence is to appeal to their being 'the same' in some way. The only thing in common between the dual theories is the empirical core, so we should take this to be what explains their epistemic equivalence. Everything else (i.e. the 'excess structure') can be discarded, since they are irrelevant differences. As such, structural realism should commit us only to the empirical core.

However, it is not clear that the dual theories are indeed epistemically equivalent. In a naive sense, they are epistemically equivalent if one takes 'epistemic' to be 'empirical' equivalence. Given the duality, i.e. formal and predictive equivalence, it is trivial that the two theories are also 'empirically' equivalent. However, I do not think such a notion of empirical equivalence *exhausts* the epistemic criteria for differentiating between scientific theories. Of course, one main desideratum for scientific theorizing is to provide

predictions, descriptions and explanations of phenomena. Beyond that, though, I contend that another desideratum of scientific theorizing is to look for ways to develop better scientific theories, be it a more unificatory theory, a more explanatory theory, and so on.

We see this in play when De Haro discusses the position/momentum duality in quantum mechanics: "this duality is usually seen as teaching us something new about the nature of reality: namely, that atoms are neither particles, nor waves. By analogy, it is to be expected that gauge/gravity dualities teach us something about the nature of spacetime and gravity" [2017, 117]. However, this is only possible *if* the two theories were not epistemically equivalent! If they were epistemically equivalent, then how could we learn anything new from one theory that we cannot already learn from another? If 'area' and 'entanglement entropy' really meant the same thing and had the same physical interpretation, how could we learn something new when we realize that area can be related (via **RT**) to quantum entanglement? Indeed, this criticism extends generally to the internal view: how can we learn anything new from a duality if the dual theories are just the 'same theory', and indeed are *uninterpreted* to begin with? We learn something new when two *different* things are related in a surprising way, *especially* when they are related to other quantities, on each side, in interesting ways; I do not see how we can learn something new when one and the same thing is related to itself.

Furthermore, the two theories are *not* epistemically equivalent when we consider the methodological concerns of physicists, who generally note that the **CFT** is well-understood, while the dual string theory of gravity is not. For example, Horowitz and Polchinski [2009] notes that we only approximately understand the gravitational theory, but the **CFT** has been developed to very precise degrees. Lin points out that:

> A dictionary is reasonably well developed in the direction of using classical gravity to study the **CFT**, but the converse problem how to organize the information in certain **CFT**'s into a theory of quantum gravity with a semi-classical limit is hardly understood at all. [2015, 11]

If both theories are equally successful by *all* epistemic criteria we have, then this situation should not appear. Rather, it seems that scientific practice is of the opinion that the two theories are, in fact, *not* epistemically equal: one is more successful than the other in terms of a variety of criteria, such as precision of calculation, ease of understanding, availability of a non-perturbative analysis, and so on. It is one reason why **AdS/CFT** is such an interesting area of research: it allows us to understand a hard-to-understand theory in terms of an easier-to-understand theory. Unless one is given arguments for why such criteria should *not* be epistemically relevant, the dual theories, I contend, are *not* epistemically equivalent.

Of course, one could assume that the *goal* or *ideal*, when we fully understand the translation manual, is to render both theories equally epistemically successful. However, this presumes that both sides *will* end up being just as easy to compute, or understand, and so on. Of course, if we do discover a more fundamental characterization of *why* the two dual theories are related by the duality as such, e.g. the sort of 'deeper' theory Rickles [2011, 2017] hopes for, then clearly we are entitled to the internal view since this 'deeper' theory will ideally explain why the dual theories, despite their apparent differences, can be seen as different facets of a single theory, just like how special relativity unified electromagnetism and made it plausible to understand both the electric and magnetic fields as facets of the 'deeper' Faraday tensor field. Right now, though,

there is no such theory in sight, making this point inadequate for supporting the internal view.

Given the foregoing, it is not clear there is epistemic equivalence: the epistemic argument does not hold. The upshot is that we are not compelled to provide an explanation for why the dual theories are epistemically equivalent to begin with (they are not), and hence we have no need to commit ourselves only to the common empirical core, *even* as structural realists, nor to think that differentiating the dual theories is meaningless.

Recall the oddity I pointed out in §3, though. The claim of physical equivalence hangs on leaving the dual theories uninterpreted, but this latter claim was itself motivated by physical equivalence. It was hoped, then, that the epistemic argument could provide independent motivation for adopting physical equivalence. Given my criticism of De Haro's additional argument, though, the circle returns, and leaves the two claims uncompelling. Hence, we should not adopt the internal view itself. Furthermore, my criticisms suggest that the dual theories are in fact *not* epistemically equivalent, and this suggests that the default stance is one where the two theories are not theoretically equivalent at all. Given the duality, the only way this can be so is to adopt the view that the dual theories are physically non-equivalent; in other words we should adopt the external view instead.

To conclude, given the dialectic set up by the interpretive fork, and the inadequacies of the internal view, I suggest that we adopt the external view instead.

5 The Way Forward

Let me end by commenting on the external view and the broader debate on whether there is emergence given a duality (§1). In §3 we have seen how the internal view precludes emergence simply because there are no two distinct theories to speak of: we merely have two ways of looking at a single theory. This in turn swiftly rules out any talk of emergence. The external view, though, does not rule out emergence quite so easily, and there is some leeway to speak of emergence since we *do* have two distinct theories which are, as Teh noted, generically *not* isomorphic to one another. However, given the formal and predictive equivalences demanded by a duality relation, a duality relation is symmetric, and so there is nothing within a duality that will formally broker the asymmetry between two theories we often associate with emergence. One way to do so, as Teh (2013) suggests, is to introduce a claim of relative fundamentality, i.e. which theory is 'more fundamental' than another, is required to break the symmetry and provide us with the required asymmetry for emergence. While the external view does not entail this, it does not rule it out either. Hence, the external view does not preclude emergence; instead, it directs attention about emergence and duality away from the interpretative fork, onto whether and how one can make claims about relative fundamentality in the context of dualities. Alas, this requires much more attention than I can afford here: I leave it for another day.

References

Dieks, D., J. van Dongen, and S. D. Haro (2015). Emergence in Holographic Scenarios for Gravity. *Studies in the History and Philosophy of Modern Physics 52*, 203–216. 10.1016/j.shpsb.2015.07.007.

Fraser, D. (2017). Formal and Physical Equivalence in Two Cases in Contemporary Quantum Physics. *Studies in the History and Philosophy of Modern Physics 59*, 30–43. 10.1016/j.shpsb.2015.07.005.

Haro, S. D. (2017). Dualities and Emergent Gravity: Gauge/Gravity Duality. *Studies in the History and Philosophy of Modern Physics 59*, 109–125. DOI: 10.1016/j.shpsb.2015.08.004.

Haro, S. D., N. Teh, and J. Butterfield (2017). Comparing Dualities and Gauge Symmetries. *Studies in the History and Philosophy of Modern Physics 59*, 68–80. DOI: 10.1016/j.shpsb.2016.03.001.

Horowitz, G. and J. Polchinski (2009). Gauge/gravity duality. In D. Oriti (Ed.), *Approaches to quantum gravity: Toward a new understanding of space time and matter*, pp. 169–186. Cambridge: Cambridge University Press. arXiv:gr-qc/0602037.

Raamsdonk, M. V. (2010). Building up spacetime with quantum entanglement. *General Relativity and Gravitation 42*(10), 2323–2329. 10.1007/s10714-010-1034-0.

Rickles, D. (2011). A Philosopher Looks at Dualities. *Studies in the History and Philosophy of Modern Physics 42*(1), 54–67. DOI: 10.1016/j.shpsb.2010.12.005.

Rickles, D. (2017). Dual Theories: 'Same but Different or 'Different but Same? *Studies in the History and Philosophy of Modern Physics 59*, 62–67. 10.1016/j.shpsb.2015.09.005.

Ryu, S. and T. Takayanagi (2006). Holographic Derivation of Entanglement Entropy from AdS/CFT. *Phys. Rev. Lett 96*(18). 10.1103/PhysRevLett.96.181602.

Seiberg, N. (2007). Emergent spacetime. In D. Gross, M. Henneaux, and A. Sevrin (Eds.), *The Quantum Structure of Space and Time*, pp. 163–178. Singapore: World Scientific. DOI: 10.1142/9789812706768_0005.

Teh, N. (2013). Holography and Emergence. *Studies in the History and Philosophy of Modern Physics 44*(3), 300–311. DOI: 10.1016/j.shpsb.2013.02.006.

Causal Discovery and the Problem of Psychological Interventions

PSA 2018, Seattle

Markus Eronen

University of Groningen

m.i.eronen@rug.nl

Abstract

Finding causes is a central goal in psychological research. In this paper, I argue that the search for psychological causes faces great obstacles, drawing from the interventionist theory of causation. First, psychological interventions are likely to be both fat-handed and soft, and there are currently no conceptual tools for making causal inferences based on such interventions. Second, holding possible confounders fixed seems to be realistically possible only at the group level, but group-level findings do not allow inferences to individual-level causal relationships. I also consider the implications of these problems, as well as possible ways forward for psychological research.

1. Introduction

A key objective in psychological research is to distinguish causal relationships from mere correlations (Kendler and Campbell 2009; Pearl 2009; Shadish and Sullivan 2012). For example, psychologists want to know whether having negative thoughts is a cause of anxiety instead of just being correlated with it: If the relationship is causal, then the two are not just spuriously hanging together, and intervening on negative thinking is actually one way of reducing anxiety in patients suffering from anxiety disorders. However, to what extent is it actually possible to find psychological causes? In this paper, I will seek an answer this question from the perspective of state-of-the-art philosophy of science.

In philosophy of science, the standard approach to causal discovery is currently interventionism, which is a very general and powerful framework that provides an account of the features of causal relationships, what distinguishes them from mere correlations, and what kind of knowledge is needed to infer them (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009, Woodward 2003, 2015b; Woodward & Hitchcock 2003). Interventionism has its roots in Directed Acyclic Graphs (DAGs), also known as causal Bayes nets, which are graphical representations of causal relationships based on conditional independence relations (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009). More recently, James Woodward has developed interventionism into a full-blown philosophical account of causation, which has become popular in philosophy and the sciences. Several authors have also argued that interventionism adequately captures the role of causal thinking and reasoning in psychological research (Campbell 2007; Kendler and Campbell 2009; Rescorla forthcoming; Woodward 2008).

Based on interventionism, I will argue in this paper that the discovery of psychological causes faces great obstacles. This is due to problems in performing psychological interventions and deriving interventionist causal knowledge from psychological data.[1] Importantly, my focus is not on the existence or possibility of psychological causation, but on the *discovery* of psychological causes, which is a topic that has so far received little attention in philosophy.[2] Although I rely on interventionism, my arguments are based on rather general principles of causal inference and reasoning in science, and will thus apply to any other theory of causation that does justice to such principles.

The focus in this paper will be on the discovery *individual-level* (or within-subject) causes, not *population-level* (or between-subjects) causes. The first refers to causal relationships that hold for a particular individual: for example, John's negative thoughts cause John's problems of concentration. The latter refers to causal relationships that obtain in the population as a whole: for example, negative thoughts cause problems of concentration in a population of university students. It widely thought that ultimate goal of causal inference is to find individual-level causes, and that a population-level causal relationship should be seen as just an average of individual-level causal relationships (Holland 1986): For example, the causal relationship between negative thoughts and problems of concentration in a population of university students is only interesting insofar as it *also* applies to at least some of the individual students in the

[1] See Eberhardt (2013; 2014) for different (and domain-independent) problems for interventionist causal discovery.
[2] There is an extensive debate on the question whether interventionism vindicates non-reductive psychological causation by providing a solution the causal exclusion problem (e.g., Baumgartner 2009, Eronen 2012, Raatikainen 2010, Woodward 2015). I will sidestep this debate here, as my focus is not on the existence of non-reductive psychological causation, but on the discovery of psychological causes, be they reducible or not.

population.[3] Thus, in this paper I will discuss population-level causal relationships only when they are relevant to discovering individual-level causes.

Importantly, the distinction between population-level and individual-level causation is different from the distinction between type and token causation, even though the two distinctions are sometimes mixed up in the philosophical literature (see also Illari & Russo 2014, ch. 5). Token causation refers to causation between two actual events, whereas type causation refers to causal relationships that hold more generally. Individual-level causes can be either type causes or token causes. An example of an individual and type causal relationship would be that John's pessimistic thoughts cause John's problems of concentration: This is a general relationship between two variables, and not a relationship between two actual events. An example of an individual and token causal relationship would be that John's pessimistic thoughts before the exam on Friday at 2 pm caused his problems of concentration in the exam. As interventionism is a type-level theory of causation, and the aim of psychological research is primarily to discover regularities, not explanations to particular events, in this paper I will only discuss the discovery of type (individual) causes.

The structure of this paper is as follows. I will start by giving a brief introduction to interventionism, and then turn to problems of interventionist causal inference in psychology: First, to problems related to psychological interventions (section 2), and then to problems arising from the requirement to "hold fixed" possible confounders (section 3). After this, I will consider the possibility of the inferring psychological causes without interventions (section 4). In the last

[3] It has been argued that population-level (between-persons) causal relationships can also be real without applying to any individual (Borsboom, Mellenbergh, and van Heerden 2003). However, also those who believe in these kind of population-level causes agree that discovering individual causes is an important goal as well.

section, I discuss ways forward and various implications that my arguments have for psychology and its philosophy.

2. Interventionism

Interventionism is a theory of causation that aims at elucidating the role of causal thinking in science, and defining a notion of causation that captures the difference between causal relationships and mere correlations (Woodward 2003). Thus, the goal of interventionism is to provide a methodologically fruitful account of causation, and *not* to reduce causation to non-causal notions or analyse the metaphysical nature of causation (Woodward 2015b). In a nutshell, interventionist causation is defined as follows:

X is a cause of Y (in variable set **V**) if and only if it is possible to *intervene* on X to change Y when all other variables (in **V**) that are not on the path from X to Y are *held fixed* to some value (Woodward 2003).

Thus, in order to establish that X is a cause of Y, we need evidence that there is some way of intervening on X that results in a change in Y, when off-path variables are held fixed.[4] Importantly, it is not necessary to actually perform an intervention: What is necessary is knowledge on what *would* happen if we *were* to make the right kind of intervention.

[4] More precisely, this is the definition for a *contributing* cause. X is a *direct* cause of Y if and only if it is possible to intervene on X to change Y when all other variables (in **V**) are held fixed to some value (Woodward 2003). Thus, the definition of a contributing cause allows there to be other variables on the causal path between X and Y, whereas the definition of a direct cause does not. This does not reflect any substantive metaphysical distinction, as the question whether X is a direct or contributing cause is relative to what variables are included in the variable set. Importantly, notion of a contributing cause is *not* relative to a variable set in any strong sense – if X is a cause of Y in some variable set, then X will be a cause of Y in all variable sets where X and Y appear (Woodward 2008b). This is because the definition of an intervention is not relativized to a variable set.

The notion of an intervention plays a fundamental role in the account, and is very specifically defined. Here is a concise description of the four conditions that an intervention has to satisfy (Woodward 2003).

Variable *I* is an intervention variable for X with respect to Y if and only if:

(I1) *I causes* the change in X;

(I2) The change in X is *entirely* due to *I* and not any other factors;

(I3) *I* is not a cause of Y, or any cause of Y that is not on the path from X to Y;

(I4) *I* is *uncorrelated* with any causes of Y that are not on the path from X to Y.

The rationale behind these conditions is that if the intervention does not satisfy them, then one is not warranted to conclude that the change in Y was (only) due to the intervention on X. Thus, in simpler terms, the intervention should be such that it changes the value of the target variable *X* in such a way that the change in *Y* is *only* due to the change in *X* and not any other influences (Woodward 2015b). For example, if the intervention is correlated with some other cause of Y, say Z, that is not on the path from X to Y (violating I4), then the change in Y may have been (partly) due to Z, and not just due to X. Following standard terminology in the literature, I will call interventions that satisfy the criteria I1-I4 *ideal* interventions. I will now go through various problems in performing ideal interventions in psychology, starting from problems related to conditions I2 and I3 (section 3), and then turn to problems related to I4 and the "holding fixed" part of the definition of causation (section 4).

3. Psychological interventions

Before discussing psychological interventions, an important distinction needs to be made: The distinction between relationships where (1) the cause is *non-psychological*, and the *effect* is psychological, and (2) where the *cause* (and possibly also the effect) is *psychological*.[5] A large proportion (perhaps the majority) of experiments in psychology involve relationships of the first kind: The intervention targets a non-psychological variable (X) such as medication vs. placebo, therapy regime vs. no therapy, or distressing vs. neutral video, and the psychological effect of the manipulation of this non-psychological variable is tracked. In other words, the putative causal relation is between a non-psychological cause variable (X) and a psychological effect variable (Y). In these cases, it is possible to do (nearly) ideal interventions on the putative cause variable (X) by ensuring that the change in X was caused (only) by the intervention, that the intervention did not change Y directly, and that it was uncorrelated with other causes of Y. It is of course far from trivial to make sure that these conditions were satisfied, but as the variables intervened upon are non-psychological, making the right kinds of interventions is in principle not more difficult than in other fields. As regards the psychological effect variable (Y), there is no need to intervene on it; it is enough to measure the change in Y (which, again, is far from trivial, but faces just the usual problems in psychological measurement, which will be discussed below). The fact that many psychological experiments involve this kind of causal relationships may have contributed to the recent optimism on the prospects of interventionist causal inference in psychology.

[5] The line between psychological and non-psychological variables is likely to be blurry. However, for the present purposes it is not crucial where exactly the line should be drawn: My arguments apply to cases where it is clear that the cause variable is psychological (such as the examples in the main text), and such cases abound in psychological research.

However, psychological research also often concerns relationships of the second kind, that is, relationships where the *cause* is psychological. This is, for example, the case when the aim is to uncover psychological mechanisms that explain cognition and behavior (e.g., Bechtel 2008, Piccinini & Craver 2011), or to find networks of causally interacting emotions or symptoms (e.g., Borsboom & Cramer 2013). The reason why these relationships are crucially different from relationships of the first kind is that now the variable intervened upon is psychological, so the conditions on interventions now have to be applied to psychological variables.

Ideal interventions on psychological variables are rarely if ever possible. One reason for this has been extensively discussed by John Campbell (2007): Psychological interventions seem to be "soft", meaning that the value of the target variable *X* is not completely determined by the intervention (Eberhardt & Scheines 2007; see also Kendler and Campbell 2009; Korb and Nyberg 2006). In other words, the intervention does not "cut off" all causal arrows ending at *X*. As a non-psychological example, when studying shopping behaviour during one month by intervening on income, an ideal intervention would fully determine the exact income that subjects have that month, whereas simply giving the subjects an *extra* 5000€ would count as a soft intervention (Eberhardt & Scheines 2007). Similarly, if we intervene on John's psychological variable *alertness* by shouting "WATCH OUT!", this does not completely cut off the causal contribution of other psychological variables that may influence John's *alertness*, but merely adds something on top of those causal contributions (Campbell 2007). As most or all interventions on psychological variables are likely to be soft, Campbell proposes that we should simply allow such soft interventions in the context of psychology. Campbell argues that these kind of interventions can still be informative and indicative of causal relationships (Campbell

2007), and this conclusion is supported by independent work on soft interventions in the causal modelling literature (e.g., Eberhardt & Scheines 2007; Korb and Nyberg 2006).

However, the problem of psychological interventions is not solved by allowing for soft interventions. There is a further, equally important reason why interventions on psychological variables are problematic: Psychological interventions typically *change several variables simultaneously*. For example, suppose we wanted to find out whether *pessimistic thoughts* cause *problems in concentration*. In order to do this, we would have to find out what would happen to *problems in concentration* if we were to intervene just on *pessimistic thoughts* without perturbing other psychological states with the intervention. However, how could we intervene on *pessimistic thoughts* without changing, for example, *depressive mood* or *feelings of guilt*? As an actual scientific example, consider a network of psychological variables that includes, among others, the items *alert*, *happy*, and *excited* (Pe et al. 2015). How could we intervene on just one of those variables without changing the others?

One reason why performing "surgical" interventions that only change one psychological variable is so difficult is that there is no straightforward way of manipulating or changing the values of psychological variables (as in, for example, electrical circuits). Interventions in psychology have to be done, for example, through verbal information (as in the example of John above) or through visual/auditory stimuli, and such manipulations are not precise enough to manipulate just one psychological variable. Also state-of-the-art neuroscientific methods such as Transcranial Magnetic Stimulation affect relatively large areas of the brain, and are not suited for intervening on specific psychological variables. Currently, and in the foreseeable future, there is no realistic

way of intervening on a psychological variable without at the same time perturbing some other psychological variables.

Thus, it is likely that most or even all psychological interventions do not just change the target variable X, but also some other variable(s) in the system. In the causal modelling literature, interventions of this kind have been dubbed *fat-handed*[6] interventions (Baumgartner and Gebharter 2016; Eberhardt & Scheines 2007; Scheines 2005). For example, an intervention on pessimistic thoughts that also immediately changes depressive mood is fat-handed. Fat-handed interventions have been recently discussed in philosophy of science, but mainly in the context of mental causation and supervenience (e.g., Baumgartner and Gebharter 2016, Romero 2015), and the fact that psychological interventions are likely to be systematically fat-handed (for reasons unrelated to supervenience) has not yet received attention.

An additional complication is that it is difficult check what a psychological intervention precisely changed, and to what extent it was fat-handed (and soft). In fields such as biology or physics there are usually several independent ways of measuring a variable: for example, temperature can be measured with mercury thermometers or radiation thermometers, and the firing rate of a neuron can be measured with microelectrodes or patch clamps. However, measurements of psychological variables, such as emotions or thoughts, are based on self-reports, and there is no further independent way of verifying that these reports are correct. Moreover, only a limited number of psychological variables can be measured at a given time point, so an intervention may always have unforeseen effects on unmeasured variables.

[6] According to Scheines (2005), this term was coined by Kevin Kelly.

Why are fat-handed interventions so problematic for interventionist causal inference? The reason becomes clear when looking at condition I3: The intervention should not change any variable Z that is on a causal pathway that leads to Y (except, of course, those variables that are on the path between X and Y). If the causal structure of the system under study is known, as well as the changes that the intervention causes, then this condition can sometimes be satisfied even the intervention was fat-handed. However, in the context of intervening on psychological variables, neither the causal structure nor the exact effects of the interventions are known. Thus, when the intervention is fat-handed, it is not known whether I3 is satisfied or not, and in many cases it is likely to be violated. In other words, we cannot assume that the intervention was an unconfounded manipulation of X with respect to Y, and cannot conclude that X is a cause of Y.

4. The Problem of "Holding Fixed"

The next problem that I will discuss is related to the last part of the definition of interventionist causation: X is a cause of Y (in variable set **V**) if and only if it is possible to intervene on X to change Y *when all other variables (in **V**) that are not on the path from X to Y are held fixed to some valu*e. The motivation for this requirement is to make sure that the change in Y is really due to the change X, and not due to some other cause of Y. To a large extent, this is just another way of stating what is already expressed in the definition of an intervention, in conditions I3 and I4: The intervention should not be confounded by any cause of Y that is not on the path between X and Y.[7] In the previous section, we saw that fat-handed interventions pose a challenge for

[7] In recent publications, Woodward often gives a shorter definition of causation that does not include the "holding fixed" part, for example: "X causes Y if and only if under some interventions on X (and possibly other variables) the value of Y changes" (Woodward 2015). This is understandable, as the definition of intervention already contains conditions I3 and I4, which effectively imply holding fixed potential causes of Y that are correlated with the intervention and are not on the path from X to Y. However, there are also good reasons why the full definition has to

satisfying this condition. However, as I will now show, it is problematic in psychology also for more general reasons.

In psychology, it is impossible to hold psychological variables fixed in any concrete way: We cannot "freeze" mental states, or ask an individual to hold her thoughts constant. Thus, the same effect has to be achieved indirectly, and the gold standard for this is Randomized Controlled Trials (RCTs) (Woodward 2003, 2008). RCTs have their origin in medicine, but are widely used in psychology and the social sciences as well (Clarke et al. 2014; Shadish, Cook and Campbell 2002; Shadish and Sullivan 2012). The basic idea of RCTs is to conduct a trial with two groups, the test group and the control group, which are as similar to one another as possible, but the test group receives the experimental manipulation and the control group does not. If the groups are large enough and the randomization is done correctly, any differences between the groups should be only due to the experimental manipulation. If everything goes well, this in effect amounts to "holding fixed" all off-path variables.

However, this methodology has an important limitation that has been overlooked in the literature on interventionism. As the effect of "holding fixed" is based on the difference between the groups as wholes, it only applies at the level of the group, and not at the level of individuals. For this reason, results of RCTs hold for the study population as a whole, but not necessarily for particular individuals in the population (cf. Borsboom 2005, Molenaar & Campbell 2009). For example, if we discover that pessimistic thoughts are causally related to problems of

include the second component as well. For example, consider a situation where we intervene on X with respect to Y, and Y changes, but this change is fully due to a change in variable Z, which is a cause of Y that is *uncorrelated* with the intervention variable. In this situation, without the "holding fixed" requirement we would falsely conclude that X is a cause of Y.

concentration in the population under study, it does not follow that this causal relationship holds in John, Mary, or any other specific individual in the population. This is related to the "fundamental problem of causal inference" (Holland 1986): Each individual in the experiment can belong to only one of the two groups (control or test group), and therefore cannot act as a "control" for herself, so only an average causal effect can be estimated. What this implies for causal inference in psychology is that when a causal relationship is discovered through an RCT, we cannot infer that this relationships holds for any specific individual in the population (see also Illari & Russo 2014, ch. 5).

This does not mean that the population-level findings based on RCTs are uninformative or useless. The point is rather that we currently have no understanding of when, to what extent and under what circumstances they also apply to the individuals in the population. This of course applies also to other fields where RCTs are used, such the biomedical sciences. Indeed, especially in the context of personalized medicine, the fact that RCTs are as such not enough to establish individual-level causal relationships has recently become a matter of discussion (e.g., de Leon 2012).

It might be tempting to simply look at the data more closely and find those individuals for whom the intervention on X actually corresponded with a change in Y. However, it would be a mistake to conclude that in those individuals the change in Y was caused by X. It might very well have been caused by some other cause of Y, as possible confounders were only held fixed at the group level, not at the individual level.[8] Thus, in RCTs possible confounders can only be held fixed at

[8] Would it be possible for a causal relationship to hold at the population level, but not for any individual in the population? Probably not, if the relationship is genuine: Weinberger (2015) has argued that there has to be at least *one* individual in the population for whom the relationships holds. However, in the context of discovery, it is

the group level, and this does not warrant causal inferences that apply to specific individuals. This is further limitation to interventionist causal inference in psychology.

5. Finding psychological causes without interventions

One possible response to the concerns raised in the previous two sections is that interventionism does not require that interventions are actually performed: As briefly mentioned in section 2, what is necessary is to know what *would* happen if we *were* to perform the right kinds of interventions. In other words, in order to establish that X is a cause of Y, it is enough to know that if we *were* to intervene on X with respect to Y (while holding off-path variables fixed), then Y *would* change. For example, it is beyond doubt that the gravitation of the moon causes the tides, even though no one has ever intervened on the gravitation of the moon to see what happens to the tides, and such an intervention would be practically impossible (Woodward 2003). Similarly, it could be argued that even though it is practically impossible to do (ideal) interventions on psychological variables, the knowledge on the effects of interventions could be derived in some other way. Let us thus consider to what extent this could be possible.

The state-of-the-art method for deriving (interventionist) causal knowledge when data on interventions is not available is *Directed Acyclic Graphs (DAGs)*, which were briefly mentioned in the introduction (see also Malinsky & Danks 2018, Pearl 2000, Spirtes, Glymour and Scheines 2000, Spirtes & Zhang 2016). Causal discovery algorithms based on DAGs take purely

possible that a causal *finding* at the population level is just an artefact of heterogeneous causal structures at the individual level, and therefore does not apply to any individual in the population.

observational data as input, and based on conditional independence relations, find the causal graph that best fits the data. In principle, these algorithms can be used for psychological data, with the aim of discovering causal relationships between psychological variables.

However, even though these algorithms do not require experimental data, they do require data from which conditional independence relations can be reliably drawn, and they (implicitly) assume that the variables that are modelled are independently and surgically manipulable (Malinsky & Danks 2018). In contrast, as should be clear from the above discussion, measurements of psychological variables typically come with a great deal of uncertainly, and it is not clear to what extent they can be independently manipulated. Moreover, causal discovery algorithms standardly assume *causal sufficiency*, that is, that there are no unmeasured common causes that could affect the causal structure (Malinsky & Danks 2018; Spirtes & Zhang 2016). The reason for this is that if two or more variables in the variable set have unmeasured common causes, then the inferences concerning the causal relationships between those variables will be either incorrect or inconclusive. However, missing common causes is likely the norm rather than the exception when it comes to psychological variables. For example, if the variable set consists of, say, 16 emotion variables, how likely is it that *all* relevant emotion variables have been included? And even if all emotion variables that are common causes to other emotion variables are included, is it plausible to assume that there are no further cognitive or biological variables that could be common causes to some of the emotion variables? As similar questions can be asked for any context involving psychological variables, causal sufficiency is a very unrealistic assumption for psychological variable sets.

For these reasons, psychological data sets are rather ill-suited for causal discovery algorithms, and these algorithms cannot be treated as reliable guides to interventionist causal knowledge in psychology. It is likely that the problems of psychological interventions discussed in the previous sections are not just practical problems in carrying out interventions, but reflect the immense complexity of the system under study (the human mind-brain), and therefore cannot be circumvented by just using non-experimental data (see, however, section 7 for a different approach).

6. Psychological interventions: A summary

To summarize, what I have argued so far is that interventionist causal inference in psychology faces several obstacles: (1) Psychological interventions are typically *both* fat-handed *and* soft: They change several variables simultaneously, and do not completely determine the value(s) of the variable(s) intervened upon. It is not known to what extent such interventions give leverage for causal inference. (2) Due to the nature psychological measurement, the degree to which a psychological intervention was soft and fat-handed, or more generally, what the intervention in fact did, is difficult to reliably estimate. (3) Holding fixed possible confounders is only possible at the population level, not at the individual level, and it is not known under what conditions population-level causal relationships also apply to individuals. (4) Causal inference based on data without interventions requires assumptions that are unrealistic for psychological variable sets. Taken together, these issues amount to a formidable challenge for finding psychological causes.[9]

[9] Baumgartner (2009, 2012, 2018) has argued that mental-to-physical supervenience makes it impossible to satisfy the Woodwardian conditions on interventions, and that if interventionism is modified to accommodate supervenience relationships (as in Woodward 2015), the result is that any causal structure with a psychological

7. Discussion

Although various metaphysical and conceptual issues related to psychological causation have been extensively discussed in philosophy of science, little attention has been paid to the *discovery* of psychological causes. In this paper, I have contributed to filling this lacuna, by discussing the search for psychological causes in the framework of the interventionist theory of causation. The upshot is that finding individual psychological causes faces daunting challenges. The problems in holding fixed confounders and performing interventions need to be taken into account when trying to establish a psychological causal relationship, or when making claims about psychological causes.

However, I do not want to argue that finding psychological causes is *impossible*, or that researchers should stop looking for psychological causes. Rather, my aim is to contribute to getting a better understanding of the limits of finding causes in psychology, and the challenges involved. This can also lead to positive insights regarding causal inference in psychology. One such insight is that more attention should be paid to *robust inference* or *triangulation*. Often when individual methods or sources of evidence are insufficient or unreliable, as is the case here, what is needed is a more holistic approach. A widespread (though not uncontroversial) idea in philosophy of science is that evidence from several independent sources can lead to a degree of confidence even if the sources are individually fallible and insufficient (Eronen 2015, Kuorikoski

cause becomes empirically indistinguishable from a corresponding structure where the psychological variable is epiphenomenal. If this reasoning is correct, it leads to a further (albeit more theoretical) problem for interventionist causal inference: Any empirical evidence for a causal relationships with a psychological cause is equally strong evidence for a corresponding epiphenomenal structure, and it is not clear which structure should be preferred and on what grounds.

& Marchionni 2017, Munafo & Smith 2017, Wimsatt 1981, 1994/2007). For example, there is no single method or source of evidence that would be individually sufficient to establish that the anthropogenic increase in carbon dioxide is the cause for the rise in global temperature, but there is so much converging evidence from many independent sources that scientists are confident that this causal relationship exists. Similarly, evidence for a psychological causal relationship could be gathered from many independent sources: Several different (soft and fat-handed) interventions involving different variables, multilevel models based on time-series data, single-case observational studies, and so on.[10] If they all point towards the same causal relationships, this may lead to a degree of confidence in the reality of that relationship. However, how this integration of evidence would exactly work, and whether it can actually lead to sufficient evidence for psychological causal relationships, are open questions.

A related point is that psychological research can also make substantive progress *without* establishing causal relationships. Often important discoveries in psychology have not been discoveries of causal relationships, but rather discoveries of robust *patterns* or *phenomena* (Haig 2012, Rozin 2001, Tabb and Schaffner 2017). Consider, for example, the celebrated discovery that people often do not reason logically when making statistical predictions, but rely on shortcuts, for example, grossly overestimating the likelihood of dying in an earthquake or terror attack (Kahneman & Tversky 1973). In other words, when we reason statistically, we often rely on heuristics that lead to biases. The discovery of this phenomenon had nothing to with methods of causal inference (Kahneman and Tversky 1973), and its significance is not captured by describing causal relationships between variables. In fact, the causal mechanisms underlying the

[10] See also Peters, Bühlmann, & Meinshausen (2016), who present a formal model for inferring causal relationships based on their stability under different kinds of (non-ideal) interventions.

heuristics and biases of reasoning are still unknown. Similar examples abound in psychology: Consider, for example, groupthink or inattentional blindness. Of course, there are likely to be causal mechanisms that give rise to these phenomena, but the phenomena are highly relevant for theory and practice even when we know little or nothing about those underlying mechanisms (which is the current situation). This, in combination with the challenges discussed in this paper, suggests that (philosophy of) psychology might benefit from reconsidering the idea that discovering causal relationships is central for making progress in psychology.

Finally, one might wonder whether the problems I have discussed here are restricted to just psychology. Indeed, I believe that the arguments I have presented are more general, and apply to any other fields where there are similar problems with soft and fat-handed interventions and controlling for confounders. There is probably a continuum, where psychology is close to one end of the continuum, and at the other end we have fields where ideal interventions can be straightforwardly performed and variables can be easily held fixed, such as engineering science. Fields such as economics and political science are probably close to where psychology is, as they also face deep problems in making (ideal) interventions and measuring their effects. Same holds for neuroscience, at least cognitive neuroscience: The problems of soft and fat-handed interventions and holding variables fixed apply just as well to brain areas as to psychological variables (see also Northcott forthcoming). Thus, appreciating the challenges I have discussed here and considering possible reactions to them could also benefit many other fields besides psychology.

To conclude, I have argued in this paper that there are several serious obstacles to the discovery of psychological causes. As it is widely assumed in both psychology and its philosophy that the discovery of causes is a central goal, these obstacles need to be explicitly discussed, taken into account, and studied further.

References

Baumgartner, M. (2013). Rendering Interventionism and Non-Reductive Physicalism Compatible. *dialectica* 67: 1-27.

Baumgartner, M. (2018). The Inherent Empirical Underdetermination of Mental Causation. *Australasian Journal of Philosophy.*

Baumgartner, M and Gebharter, A. (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science* 67: 731-756.

Borsboom, Dennny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press

Borsboom, Denny and Anqelique O. Cramer. 2013. "Network analysis: an integrative approach to the structure of psychopathology." *Annual review of clinical psychology* 9: 91-121.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203.

Campbell, John. 2007. "An interventionist approach to causation in psychology." In: A. Gopnik & L. Schulz (eds.) *Causal Learning. Psychology, Philosophy, and Computation*. Oxford: Oxford University Press, 58–66.

Chirimuuta, Mazviita. Forthcoming. “Explanation in Computational Neuroscience: Causal and Non-causal.” *British Journal for the Philosophy of Science*. DOI:https://doi.org/10.1093/bjps/axw034

Clarke, B., Gillies, D., Illari, P, Russo, F., & Williamson, J. 2014. “Mechanisms and the evidence hierarchy.” *Topoi* 33: 339-360.

de Leon, J. (2012). Evidence-based medicine versus personalized medicine: are they enemies? *Journal of clinical psychopharmacology*, *32*(2), 153-164.

Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science*, *80*(5), 684-696.

Eberhardt, F. (2014). Direct causes and the trouble with soft interventions. *Erkenntnis*, *79*(4), 755-777.

Eberhardt, Frederick and Richard Scheines. 2007. “Interventions and causal inference.” *Philosophy of Science* 74: 981–995.

Eronen, Markus. Forthcoming. “Interventionism for the Intentional Stance: True Believers and Their Brains.” *Topoi*.

Hamaker, Ellen L. 2011. “Why researchers should think **“**within-person.**””** In M. R. Mehl, & T. A. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Kahneman, Daniel and Amos Tversky. 1973. “On the psychology of prediction.” *Psychological Review* 80: 237-251.

Kendler, Kenneth S. and John Campbell. 2009. Interventionist causal models in psychiatry: repositioning the mind-body problem. *Psychological Medicine* 39: 881-887.

Korb, K. B., & Nyberg, E. 2006. "The power of intervention." *Minds and Machines* 16*:* 289-302.

Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science, 83*, 227-247.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470.

Menzies, Peter. 2008. "The exclusion problem, the determination relation, and contrastive causation." In J. Hohwy & J. Kallestrup (Eds.) *Being Reduced* (pp. 196-217). Oxford: Oxford University Press.

Molenaar, Peter and Cynthia Campbell. 2009. "The new person-specific paradigm in psychology." *Current Directions in Psychological Science* 18: 112-117.

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature* 553, 399-401

Northcott, R. (forthcoming). Free will is not a testable hypothesis. *Erkenntnis*.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... & Kuppens, P. 2015. "Emotion-network density in major depressive disorder." *Clinical Psychological Science*, *3*(2), 292-300.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, Judea. 2009. "Causal inference in statistics: An overview." *Statistics surveys* 3: 96-146.

Pearl, Judea. 2014. "Comment: understanding simpson's paradox." *The American Statistician* 68: 8-13.

Peters, J. , Bühlmann, P. and Meinshausen, N. (2016), Causal inference by using invariant prediction: identification and confidence intervals. J. R. Stat. Soc. B, 78: 947-1012. doi:10.1111/rssb.12167

Rescorla, Michael. Forthcoming. "An interventionist approach to psychological explanation." *Synthese*.

Reutlinger, Alexander and Juha Saatsi (eds.). 2017. *Explanation Beyond Causation*. Oxford: Oxford University Press.

Romero, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese*, *192*(11), 3731-3755.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2-14.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science*, *72*(5), 927-940.

Shadish W. R., Cook T. D. and Campbell D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin; Boston.

Shadish, W. R., & Sullivan, K. J. 2012. "Theories of causation in psychological science." In H. Cooper et al. (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 23-52). Washington, DC: American Psychological Association.

Shapiro, Lawrence. 2010. "Lessons from causal exclusion." *Philosophy and Phenomenological Research*, 81, 594-604.

Shapiro, Lawrence. 2012. "Mental manipulations and the problem of causal exclusion." *Australasian Journal of Philosophy*, 90, 507–524.

Shapiro, Lawrence and Elliott Sober. 2007. “Epiphenomenalism: the dos and the don’ts.” In G. Wolters & P. Machamer (Eds.) *Thinking about causes: from Greek philosophy to modern physics* (pp. 235–264). Pittsburgh, PA: University of Pittsburgh Press.

Spirtes, Peter, Glymour, Clark and Richard Scheines. 2000. *Causation, prediction, and search*. New York: Springer.

Tabb, K., & Schaffner, K. F. (2017). Causal pathways, random walks and tortuous paths: Moving from the descriptive to the etiological in psychiatry. In: Kendler, K. S., & Parnas, J. (Eds.) *Philosophical Issues in Psychiatry IV: Nosology* (pp. 342-360). Oxford: Oxford University Press.

Weinberger, Naftali. 2015. “If intelligence is a cause, it is a within-subjects cause.” *Theory & Psychology*, 25(3), 346-361.

Woodward, James. 2003. *Making things happen. A theory of causal explanation.* Oxford: Oxford University Press.

Woodward, James. 2008. “Mental causation and neural mechanisms.” In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: new essays on reduction, explanation, and causation*. Oxford: Oxford University Press: 218-262

Woodward, James. 2015a. “Interventionism and causal exclusion.” *Philosophy and Phenomenological Research* 91, 303-347.

Woodward, James. 2015b. “Methodology, ontology, and interventionism.” *Synthese* 192, 3577-3599.

Woodward, James & Christopher Hitchcock. 2003. “Explanatory Generalizations, Part I: A Counterfactual Account.” *Noûs* 37(1): 1–24.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

Why Replication is Overrated

Current debates about the replication crisis in psychology take it for granted that direct replication is valuable and focus their attention on questionable research practices in regard to statistical analyses. This paper takes a broader look at the notion of replication as such. It is argued that all experimentation/replication involves individuation judgments and that research in experimental psychology frequently turns on probing the adequacy of such judgments. In this vein, I highlight the ubiquity of conceptual and material questions in research, and I argue that replication is not as central to psychological research as it is sometimes taken to be.

1. Introduction: **The "Replication Crisis"**

In the current debate about replicability in psychology, we can distinguish between (1) the question of why not more replication studies are done (e.g., Romero 2017) and (2) the question of why a significant portion (more than 60%) of studies, when they *are* done, fail to replicate (I take this number from the Open Science Collaboration, 2015). Debates about these questions have been dominated by two assumptions, namely, first, that it is in general desirable that scientists conduct replication studies that come as close as possible to the original, and second, that the low replication rate can often be attributed to statistical problems with many initial studies, sometimes referred to as "p-hacking" and "data-massaging."[1]

[1] An important player in this regard is the statistician Andrew Gelman who has been using his blog as a public platform to debate methodological problems with mainstream social psychology (http://andrewgelman.com/).

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

I do not wish to question that close (or "direct") replications can sometimes be epistemically fruitful. Nor do I wish to question the finding that are severe problems in the statistical analyses of many psychological experiments. However, I contend that the focus on formal problems in data analyses has come at the expense of questions about the notion of *replication* as such. In this paper I hope to remedy this situation, highlighting in particular the implications of the fact that psychological experiments in general are infused with conceptual and material presuppositions. I will argue that once we gain a better understanding of what this entails with respect to replication, we get a deeper appreciation of philosophical issues that arise in the investigative practices of psychology. Among other things, I will show that replication is not as central to these practices as it is often made out to be.

The paper has three parts. In part 1 I will briefly review some philosophical arguments as to why there can be no exact replications and, hence, why attempts to replicate always involve individuation judgments. Part 2 will address a distinction that is currently being debated in the literature, i.e., that between direct and conceptual replication, highlighting problems and limitations of both. Part 3, finally, will argue that a significant part of experimental research in psychology is geared toward exploring the shape of specific phenomena or effects, and that the type of experimentation we encounter there is not well described as either direct or conceptual replication.

2. The Replication Crisis and the Ineliminability of Concepts

When scientists and philosophers talk about successfully replicating an experiment, they typically mean that they performed the same experimental operations/interventions. But what does it mean to perform "the same" operations as the ones performed by a previous experiment? With regard to this question, I take it to be trivially true that two experiments cannot be identical: At the very least, the time variable will differ. Replication can therefore at best aim for *similarity* (Shavit & Ellison 2017), as is also recognized by some authors in psychology. In this vein, Lynch et al (2015) write that "[e]xact

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

replication is impossible" (Lynch et al 2015, 2), arguing that at most advocates of direct replication can aim for is to get "as close as possible," i.e., to conduct an experiment that is similar to the previous one. In the literature, such experiments are also referred to as "direct replications." (e.g., Pashler & Harris 2912).[2]

The notion of similarity is, of course, also notoriously problematic (e.g., Goodman 1955), since any assertion of similarity between A and B has to specify with regard to what they are similar. In the context of experimentation, the relevant kinds of specifications already presuppose conceptual and material assumptions, many of which are not explicated, about the kinds of factors one is going to treat as relevant to the subject matter (see also Collins 1985, chapter 2). Such conceptual decisions will inform what one takes to be the "experimental result" down the line (Feest 2016). For example, If I am interested in whether listening to Mozart has a positive effect on children's IQ, I will design an experiment, which involves a piece by Mozart as the independent variable and the result of a standardized IQ-test at a later point. Now if I get an effect, and if I call it a Mozart effect, I am thereby assuming that the piece of music I used was causally responsible *qua being a piece by Mozart*. Moreover, when I claim that it's an effect on intelligence, I am assuming that the test I used at the end of the experiment *in fact measured intelligence*. These judgments rely on conceptual assumptions already built into the experiment qua choice of independent and dependent variables. In addition, I need *material assumptions* to the effect that potentially confounding variables have been controlled for. I take this example to show that whenever we investigate an effect *under a description*, we cannot avoid making conceptual assumptions when determining whether an experiment has succeeded or failed. This goes for original experiments as well as for replications.

[2] Both advocates and critics of direct replication sometimes contrast such replications with "conceptual" replications" (Lynch et al 2015). We will return to this distinction below.

One obvious rejoinder to this claim might be to say that replication attempts need not investigate effects under a description. They might simply imitate what the original experiment did, with no particular commitment to what is being manipulated or measured. But even if direct replications need not explicitly replicate effects under a description, I argue that they nonetheless have to make what Lena Soler calls "individuation judgments" (Soler 2011). For example, the judgment that experiment 2 is relevantly similar to experiment 1 involves the judgment that experiment 2 does not introduce any confounding factors that were absent in experiment 1. However, such judgments have to rely on some assumptions about what is relevant and what is irrelevant to the experiment, where these assumptions are often unstated auxiliaries. For example, I may (correctly or incorrectly) tacitly assume that temperature in the lab is irrelevant and hence ignore this variable in my replication attempt.

It is important to recognize that the individuation judgments made in experiments have a high degree of epistemic uncertainty. Specifically, I want to highlight what I call the problem of "conceptual scope," which arises from the question of how the respective independent and dependent variables are described. Take, for example, the above case where I play a specific piece by Mozart in a major key at a fast pace. A lot hangs on what I take to be the relevant feature of this stimulus: the fact that it's a piece by Mozart, the fact that it's in a major key, the fact that it's fast? etc. Depending on how I describe the stimulus, I might have different intuitions about possible confounders to pay attention to. For example, if I take the fact that a piece is by Mozart as the relevant feature of the independent variable, I might control for familiarity with Mozart. If I take the relevant feature to be the key, I might control for mood. Crucially, even though scientists make decisions on the basis of (implicit or explicit) assumptions about conceptual scope, their epistemic situation is typically such that they don't know what is the "correct" scope. This highlights a feature of psychological experiments that is rarely discussed in the literature about the replication crisis, i.e., the deep epistemic uncertainty and conceptual openness of much

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

research. This concerns both the initial and the replication study. Thus, concepts are ineliminable in experimental research, while at the same time being highly indeterminate.

3. Is the dichotomy between direct and less direct replication pragmatically useful?

One way of paraphrasing what was said above is that all experiments involve individuation judgments and that this concerns both original and replication studies. While this serves as a warning against a naïve reliance on direct (qua non-conceptual) replication, it might be objected that direct replications nonetheless make unique epistemic contributions. This is indeed claimed by advocates of both direct and less direct (="conceptual") replication alike. I will now evaluate claims that have aligned the distinction between direct and "conceptual" with some relevant distinctions in scientific practice, such as that between the aim of establishing the existence of a phenomenon and that of generalizing from such an existence claim on the one and that between reliability and validity on the other. I will argue that while these distinctions are heuristically useful, but on closer inspection bring to the fore exactly the epistemological issues just discussed.

3.1 Existence vs. Generalizability

Many scientists take it as given that there cannot be two identical experiments, but nonetheless argue that there is significant epistemic merit in trying to get *close enough*., i.e., to conduct direct replications. In turn, the notion of a direct replication is frequently contrasted with that of a "conceptual" replication. In a nutshell, direct replications essentially try to redo "the same" experiment (or at least something very close), whereas the conceptual replications try to operationalize the same question or concept/effect in a different way. The advantage of direct replications, as viewed by its advocates, is that by being able to redo an experiment faithfully and to create the same effect, one can show that the effect was real: "Exact and very close replications establish the basic existence and stability of a

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

phenomenon by falsifying the (null) hypothesis that observations simply reflect random noise" (LeBel et al, forthcoming, 7).

Advocates of conceptual replication don't deny this advantage of close replications, but hold that we want more than to establish that a given effect – created under very specific experimental conditions – is real. We want to know whether our findings about it can be generalized to: "When the goal is generalization, we argue that 'imperfect' conceptual replications that stretch the domain of research may be more useful" (Lynch et al 2015, 2). From a strictly Popperian perspective, the idea that non-falsification of the hypothesis of random error can provide proof of stability and existence is questionable , of course. But even if we abandon Popperian ideology here and take the falsification of H0 (that the initial effect was due to random error) to point to the truth of H1 (that there is a stable effect), the question is how to describe the effect. In other words, when claiming to have confirmed an effect, we have to say *what kind of effect* it is. And there we face the following dilemma:

a) Either we describe the effect as highly specific to very local experimental circumstances, involving the choice of a specific independent variable, delivered in a specific way etc.
b) Or we describe it in slightly broader terms, e.g., as a Mozart effect.

Advocates of direct replication might indeed endorse something like a), thereby exhibiting the kind of caution that motivated early operationists, in that no claim is made beyond the confines of a specific experiment. If, on the other hand, psychologists endorsed a description such as b), they would immediately run into the question of conceptual scope, i.e., the question *under what description* the independent variable can be said to have caused an effect. I argue that no amount of direct replication can answer this question, and hence, even if direct replication can confirm the existence of an effect, it cannot say what kind of effect. By asserting this, I am not saying that it's never useful to do a direct replication. My claim is merely that it will tell us relatively little. More pointedly: Direct replication can (perhaps) provide evidence for the existence of something, but it cannot say *existence of what*. Rolf

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

Zwaan makes a similar point when he states that "replication studies "tell us about the reliability of those findings. They don't tell us much about their validity." (Zwaan 2013).

In a similar vein, I argue that direct replication, with its narrow focus on ruling out random error, is epistemically unproductive, because it has nothing to say about *systematic error*. Systematic error arises if one erroneously attributes an effect to a specific feature of the experiment, when it is in fact due to another feature of the experiment. This can include, but is not limited to, the above-mentioned problem of conceptual scope. Fiedler et al. (2012) make a similar point when they argue that a narrow focus on falsification (with the aim of avoiding false positives) can be detrimental to the research process. Differently put, by privileging direct replication, we are not in a position to inquire about the kind of effect in question. This question, I argue, is best addressed by paying close attention to the possibility of systematic error, and hence by doing conceptual work. In other words, experimentally probing into systematic errors of conceptual scope is a valuable and productive part of the research process as it enables scientists to gradually explore what kind of effect (if any) they are looking at.[3]

3.2 Generality

I have argued that (a) scientists typically produce effects under a description and (b) that it can be epistemically productive to probe the scope of the description and to investigate the possibility of systematic error with regard to experiments that draw on such descriptions. It is epistemically productive, because it forces scientists to explore the nature and boundaries of the effect they are investigating. With this I have argued against a narrow focus on direct replication and I have cautioned against overstating the epistemic merits of such replication. But when we are concerned with effects

[3] I take this to be a contribution to arguments that philosophers of experimentation have made for a long time; e.g., Mayo 1996.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

under a description, we are confronted with questions about the adequacy of the description. It is this question that advocates of "conceptual replication" claim to be able to address when they emphasize that their approach can deliver generality (over mere existence).

We have to distinguish between two notions of generality, namely (a) what kinds of descriptions one can generalize or infer to *within the experiment*, and (b) does the effect in question hold *outside the lab* (see Feest & Steinle 2016). These types of generality are also sometimes referred to as internal vs. external validity, respectively (Campbell & Stanley 1966; Guala 2012), where the former refers to the quality of inferences within an experiment and the latter refers to the quality of inferences from a lab to the world. The notion of generalizability raises questions about two kinds of validity. My focus here will be on internal validity, i.e., with the question of whether the effect generated in an experiment really exists as described by the scientist.[4]

Internal validity can fail to hold because of epistemic uncertainties regarding confounding variables both internal and external to experimental subjects. For example, prior musical training might make a difference to how one responds to Mozart music, but the experimenter may not have taken this into consideration in their design. But internal validity can also fail to hold is by virtue of what I have referred to as the problem of conceptual scope (for example, we may refer to the effect as a Mozart effect when it is in fact a Major-key effect). Effectively, when I treat a major-key effect as a Mozart effect, I have misidentified the relevant causal feature of the stimulus. In turn, this means that I will neglect to control for major/minor key as I will regard this as irrelevant, which can result in systematic errors. In both cases, scientists can go wrong in their individuation judgment. What is at stake is not whether there is an effect, but what kind of effect it is. Now, given that those kinds of problems can

[4] In this respect I differ from some advocates of conceptual replication, who have highlighted external validity as a desideratum (E.g., Lynch 1982, 3/4).

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

occur, we turn to the question of whether "conceptual replication" has an answer. I will now argue that it does not.

To explain this, let me return to the above characterization of conceptual replication, according to which such replication consists in repeating an experiment, using different operationalizations of the same construct. For example, a conceptual replication of an experiment about the Mozart effect might operationalize the concept Mozart effect differently by using a different piece of Mozart music and/or a different measure of spatial reasoning. But there is a major caveat here: If I want to compare the results of two experiments that operationalized the same construct differently, I already have to presuppose that both operationalizations in fact have the same conceptual scope, i.e., that they in fact individuate the same effect. But this would be begging the question, since after all – given the epistemic uncertainty and conceptual openness highlighted above – that's precisely what's at issue. Differently put, experiment 2 might or might not achieve the same result as experiment 1, but the reason for this would be underdetermined by the experimental data. Thus, the problem of conceptual scope prevents us from being able to say whether we have succeeded in our conceptual replication.

Given the uncertainties as to whether one has in fact succeeded in conceptually replicating a given experiment, I am weary of the language of replication here. If anything, I would argue that the method in question should be regarded as a research strategy that is aimed at helping to demarcate and explore the very subject matter under investigation. But as I will argue now, this is perhaps better described as exploration, not as replication.

4. Putting Replication in its Proper Place

The conclusion of the previous paragraphs seems pretty bleak: Direct replication is either extremely narrow in what it can deliver or it runs into the joint problems of confounders and conceptual scope. Conceptual replication, on the other hand, cannot come to the rescue, because it also runs into the

exact same problems. Should we then throw up our hands and conclude that since ultimately neither direct nor conceptual replication are possible the crisis of replication is much more severe than we previously thought? This would be the wrong conclusion, however. This would only follow if replication was in fact as central to research as it is sometimes taken to be. I claim that it is not. My argument for these claims has three parts. The first part holds that exploring (the possibility of) systematic errors is an important part of the investigative process, which is not well described as replication. Second, if we take seriously this process of exploring and delineating the relevant phenomena, we find that there is indeed a great deal of uncertainty in psychological research, but this, in and of itself, does not necessarily constitute a crisis. Lastly, while it is fair to say that there is a crisis of confidence in current psychology, it is not well described as a replication crisis.

Let me begin with the first point. I have argued that direct replication (even where it is successful) is of limited value, because it can at most rule out random error, but completely fails to be able to address systematic error. But if we appreciate (as I have argued we should) that direct replication inevitably involves individuation judgments, it is obvious that there is always a danger of systematic error, because I have to assume that all confounding variables have been controlled for. One important class of confounders follows from what I have referred to as the problem of conceptual scope, i.e., the difficulty of correctly describing both the independent variable responsible for a given effect and the dependent variable.[5] Epistemically productive experimental work, I claim, therefore needs to focus on systematic errors, specifically those brought about by unstated auxiliary assumptions.

Indeed, if we look at the story of the Mozart effect, we find that this is exactly what happened. This example also nicely illustrates my claim about the conceptual openness and epistemic uncertainty

[5] My focus here has been mainly on the former. But of course the problem of conceptual scope concerns both.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

in many areas of experimental psychology. The Mozart effect was first posited by Rauscher and colleagues (Rauscher et a. 1993). It can now be regarded as largely debunked. While it is true that several people tried (and failed) to replicate the effect (e.g., Newman et al. 1995; Steele 1997), it is important to look at the details here. It is not the case that the effect was simply abandoned for lack of replicability. Rather, when we look at the back and forth between Rauscher and her critics, we find that the discussion turned on the choices and interpretations of independent and dependent variables. In this vein, Newman et al (1995) and Steele (1997) used different dependent variables, prompting Rauscher to argue that her effect was more narrowly confined to the kind of spatial reasoning measured by the Stanford-Binet. I suggest that we interpret this case as one where Rauscher was forced to confront (and retract) an unstated auxiliary assumption of her initial study, namely that the spatial reasoning subtest of the Stanford-Binet (which she had used as her dependent variable), was representative of spatial reasoning more generally. Likewise, her choice of the Mozart's Sonata for Two Pianos in D-major as the independent variable was put under considerable pressure by critics, who suggested that the relevant feature of the independent variable was not that it was a piece by Mozart, but that it was up-beat and put subjects in a good mood (Chabris 1999). My point here is that the debates surrounding the Mozart effect are best described as conceptual work, exploring consequences of possible errors that might have arisen from the problem of conceptual scope. At issue, I claim, was not primarily whether Rauscher really found an effect, but rather what was the scope of the effect.

I argue that this is a typical case. Rather than, or in addition to, attempting to conduct direct replications of previous experiments, researchers critically probed some hidden assumptions built into the design and interpretation of the initial experiment. My point here is both descriptive and normative. Thus, I argue that this is a productive way to proceed. However, I claim that it is not well described as replication, let alone conceptual replication. Rather, what we see here is a case in which scientists explore the empirical contours of a purported effect in the face of a high degree of epistemic

uncertainty and conceptual openness, and this is precisely why the case is not well described as employing conceptual replication. The reason for this is quite simple: For a conceptual replication to occur, one needs to already be in the possession of some well-formed concepts, such that they can be operationalized in different ways. It also presupposes that in general the domain is well-understood, such that operationalizations can be implemented and confounding variables can be controlled. But this completely misses the point that researchers often investigate effects precisely because they don't have a good understanding (and hence concept) of what it is.

Therefore I argue that while direct replication can only contribute a very small part to the research process, conceptual replication cannot make up for the shortcomings of direct replication. Instead, productive research should (and frequently does) proceed by exploring, and experimentally testing, hypotheses about possible systematic errors in experiment. Such research, I suggest, can contribute to conceptual development by helping to explore and fine-tune the shape and scope of proposed or existing concepts. The fact that this is riddled with problems does not in and of itself constitute a crisis, let alone a replication crisis.

5. Conclusion

The upshot of the above is that when we talk about the importance of replication, we need to be clear on what we mean by replication and why it is so important, precisely.

In this paper I have argued that if by replication we mean either "direct" or "conceptual" replication, we need to first of all be clear that direct replications are not non-conceptual. I then turned to some alleged epistemic merits of direct replication, for example that they can establish the existence of effects or the reliability of procedures that detect effects. I argued that insofar as such replications involve concepts, they run (among other things) into the problem of conceptual scope, i.e., the difficulty of determining, on the basis of independent and dependent variables of experiments what precisely is

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

the scope of the effect one is trying to replicate. I highlighted that this is a real and pernicious problem in experimental research in psychology, due to the high degree of epistemic uncertainty and conceptual openness of many fields of research.

While my emphasis of the conceptual nature of replication may suggest that I would be more favorably inclined toward conceptual replication, I have argued that conceptual replication runs into the same problems, and for similar reasons: The very judgement that one has successfully performed a conceptual replication of a previous experiment presupposes what is ultimately the aim of the research, namely to arrive at a robust understanding of the relevant area of research. This, I argue that since conceptual replication presupposes a relatively good grasp of the relevant concepts, it is begging the question, and I suggested instead that researchers (should) engage in a process of specifically investigating possible systematic errors in original studies as a means to develop the relevant concepts. This process is not best described as one of replication, however. Summing up, then, I conclude that in general, replications are less useful and important than is widely assumed – at least in the kind of psychological research I have focused on in this paper.

Now, in conclusion let me return to the notion of a crisis in psychology as it is currently discussed in the literature. Obviously, I do not mean to deny that there is a crisis of confidence in (social) psychology (Earp & Trafimov 2015) as well as in other areas of study. However, based on the analysis provided in this paper, I argue that this crisis is not well described as a crisis of replication. Rather, it seems to be to a large degree a crisis that turns on questionable research practices with regard to the use of statistical methods in psychology (see Gelman & Loken 2014). While acknowledging the valuable philosophical and scientific work that is being done in this area, I suggest that a broader focus on the notion of replication provides us with a deeper appreciation of the conceptual dynamics characteristic of experimental practice.

REFERENCES

Campbell, D. T., and Stanley, J. C. (1966), Experimental and Quasi-Experimental Designs for Research (Chicago: Rand McNally).

Chabris, C. (1999): Prelude or requiem for the 'Mozart Effect?' "Scientific Correspondence", Nature, 4000, 826.

Collins, H. (1985). Changing order. Replication and induction in scientific practice. Chicago and London: The University of Chicago Press.

Earp, Brian & Trafimow, David (2015): "Replication, falsification, and the crisis of confidence in social psychology." Front. Psychol, 19 May 2015 | https://doi.org/10.3389/fpsyg.2015.00621

Feest, U., 2016, "The Experimenters' Regress Reconsidered: Tacit Knowledge, Skepticism, and the Dynamics of Knowledge Generation". Studies in History and Philosophy of Science, Part A 58 34-45.

Feest, U. & Steinle, F., 2016, "Experiment." In P. Humphreys (Ed.): Oxford Handbook of Philosophy of Science. Oxford University Press, 274–295.

Fiedler, K.; Kutzner, F. & Krueger, J. (2012): „The Long Way from alpha-error control to validity proper: Problems with a short-sighted false-positive debate." Perspectives on Psychological Science 7(6), 661-669

Gelman, Andrew & Loken, Eric (2014): The Statistical Crisis in Science. Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up. American Scientist 102 (6) 460-464. DOI: 10.1511/2014.111.460

Goodman, Nelson (1983/1955): Fact. Fiction and Forecast. Harvard University Press; 4 Revised edition edition

Guala, F. (2012), "Philosophy of Experimental Economics." In U. Mäki (ed.), Handbook of the philosophy of science. Vol. 13: Philosophy of Economics (Boston: Elsevier/Academic Press), 597–640

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

LeBel, E.P.; Berger, D., Campbell, L.; Loving, T. (2017): "Falisifiability is not Optional." Journal of Personality and Social Psychology (forthcoming)

Lynch, J. (1982): "On the External Validity of Experiments in Consumer Research. Journal of Consumer Research 9, 225-239. (December)

Lynch, J.; Bradlow, E.; Huber, J.; Lehmann, D. (2015): "Reflections on the replication corner: In praise of conceptual replication." IJRM ???

Mayo, Deborah (1996): Error and the Growth of Experimental Knowledge. University of Chicago Press.

Newman, J., Rosenbach, J., Burns, K.; Latimer, B., Matocha, H., Vogt, E. (1995: An experimental test of the 'Mozart Effect': Does listening to Mozart improve spatial ability? Perceptual and Motor Skills, 81, 1379-1387.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349

Pashler, Harold & Harris, Christine (2012): "Is the Replication Crisis Overblown?" Perspectives on Psychological Science 7(6), 531-536.

Rauscher, F., Shaw, G.; Ky, K. (1993). Music and spatial task performance. Nature ,365, 611.

Romero, Felipe (2017): "Novelty vs. Replicability. Virtues and Vices in the Reward System of Science." Philosophy of Science.

Shavit, Ayelet & Ellison, Aaron (eds.) (2017): Stepping in the Same River Twice. Replication in Biological Research. Yale University Press

Soler, Lena (2011): "Tacit Elements of Experimental Practices: analytical tools and epistemological consequences." European Journal for Philosophy of Science 1, 393-433.

Steele, K., (2000). Arousal and mood factors in the 'Mozart effect'. Perceptual and Motor Skills, 91, 188-190.

Zwaan, Rolf (2013): "How Valid are our Replication Attempts?"

https://rolfzwaan.blogspot.de/2013/06/how-valid-are-our-replication-attempts.html

Author: Paul L. Franco, UW-Seattle, Department of Philosophy

Contact: pfranco@uw.edu

Title: Speech Act Theory and the Multiple Aims of Science

Abstract: I draw upon speech act theory to understand the speech acts appropriate to the multiple aims of scientific practice and the role of nonepistemic values in evaluating speech acts made relative to those aims. First, I look at work that distinguishes explaining from describing within scientific practices. I then argue speech act theory provides a framework to make sense of how explaining, describing, and other acts have different felicity conditions. Finally, I argue that if explaining aims to convey understanding to particular audiences rather than describe literally across all contexts, then evaluating explanatory acts directed to the public or policymakers involves asking nonepistemic questions.

(Accepted with minor revisions to the PSA 2018 proceedings issue of Philosophy of Science | Revisions not yet made; final version due January 2019)

1. Introduction

Hasok Chang "[complains] about…our [i.e., philosophers of science] habit of focusing on descriptive statements that are either products or presuppositions of scientific work, and our commitment to solving problems by investigating the logical relationships between these statements" (2014, 67–8). He argues philosophers of science should adopt "a change of focus from propositions to actions" (67). Chang suggests, "When we do pay attention to words, it would be better to remember to think of 'how to do things with words', to recall J. L. Austin's (1962) famous phrase" (68).

In this paper, I take Chang's suggestion and argue that attending to Austin's account of the things we do with words can help us understand the multiple goals of scientific practices, the speech acts appropriate to those goals, and the roles of nonepistemic values in evaluating speech acts made relative to those aims. In §2, I give an overview of a few philosophers of science working on explanation who have shifted focus from propositions to explaining.[1] I also briefly relate this work to a few themes in speech act theory. In §3, I give more details of Austin's framework to highlight ways of evaluating speech acts beyond truth and falsity. In §4, I explore the multiple goals of scientific practice, especially goals related to conveying understanding to the general public and policymakers, and the speech acts appropriate to those goals.

2. The things scientists do with words

2.1 Explaining

Consider some recent and not-so-recent work on scientific explanation. Andrea Woody's defense of a functional perspective on explanation aims to motivate "a shift in focus away from explanations, as achievements, toward explaining, as a coordinated activity of communities" (2015, 80). In a similar spirit, Angela Potochnik argues that when looking at explanation, "sidelining the communicative purposes to which explanations are put is a mistake" (2016, 724). She emphasizes that explaining is a communicative act involving a speaker and audience made against a background that shapes the explanations offered. In so

[1] I make no claims Chang influenced the work I canvas.

arguing, Potochnik deliberately recalls Peter Achinstein's claim, "Explaining is an illocutionary act," i.e., a speech act uttered by a speaker with a certain force and for a certain point (1977, 1).

These accounts share in common an emphasis on the importance of the aims of the speaker and audience, and thus the context of utterance in evaluating, to borrow terminology from Austin, the felicity conditions of explanatory speech acts. In particular, we might focus on the aims of the speaker and their audience in requesting and giving explanations, the time and location of an explaining speech act, and, following Woody, "what role(s) [explanations] might play in practice" (2015, 81). In focusing on the explaining act rather than the supposedly stable propositional content of an act of explanation, our attention is drawn to dimensions of evaluation beyond truth and falsity.

On this last point, Nancy Cartwright argues that the functions of a scientific theory to "tell us…what is true in nature, and how we are to explain it…are entirely different functions" (1980, 159). *Ceteris paribus* laws used in scientific theories are literally false, but still do explanatory work. One way to understand Cartwright's claim is that the speech act of describing the world truly and the speech act of explaining come apart from one another. In coming apart from one another and fulfilling different aims within scientific practice, descriptive and explanatory speech acts have different felicity conditions. For example, Potochnik (2016) examines the ways in which explaining increases understanding. But, Potochnik argues, what gets explained depends on a speaker's and audience's interests, and an explaining act's success in generating understanding depends on the cognitive resources of the audience. As such, to evaluate any given communicative act of explaining requires attending to the epistemic and nonepistemic interests of speakers and audiences that form the background against which explanations are offered. This means evaluating explanatory speech acts solely in terms of truth or falsity is inapt.

2.2 Multiple aims and the true/false fetish

I do not think this focus on acts and away from the truth or falsity of descriptive statements is unique to philosophers of science interested in explanation. We see a similar shift in work on the so-called aims approach to values in science (e.g., Elliott and McKaughan 2014;

Intemann 2015). The aims approach shares in common with work on explaining a recognition that scientific practice aims at more than describing the world truly or falsely. Further, if some of those aims include things like making timely policy recommendations for decision makers or increasing public understanding of science, there is a role for nonepistemic values in parts of scientific practice. As Kevin Elliott and Daniel McKaughan put this point, "representations can be evaluated not only on the basis of the relations that they bear to the world but also in connection with the various uses to which they are put" (2014, 3).

Why look to speech act theory to flesh out this picture about the multiple aims of scientific practice and their relationship to nonepistemic values? In part because speech act theory makes sense of the different uses to which one and the same sentence might be put depending on the aims of the speaker and audience and the context of utterance. In doing so, I think Austin is right that we can "play Old Harry with two fetishes…(1) the true/false fetish, (2) the value/fact fetish" (1962, 150). Austin was mainly content to play Old Harry with these fetishes to free philosophers from the grip of the so-called descriptive fallacy: the view "that the sole business, the sole interesting business, of any utterance…is to be true or at least false" (1970, 233). But I also think that in combating the descriptive fallacy and the true/false and fact/value fetishes, speech act theory motivates a constructive shift from the truth or falsity of descriptive statements to the things we do with words.

Take Austin's claim that evaluating apparently descriptive speech acts like "'France is hexagonal,'" involves nonepistemic questions about who is uttering the statement, in what context, and with what "intents and purposes" (1962, 142). Rather than concluding the sentence is false and leaving it at that, Austin points out the different speech acts one can use such a sentence to perform, e.g., stating or interpreting or estimating. In determining the use the sentence is put to—with the help of context and by inquiring after the interests of the speaker and their audience—we might realize, irrespective of the sentence's literal truth or falsity, "It is good enough for a top-ranking general, perhaps, but not for a geographer" (142). In other words, it serves the aims of the general, which, unlike the aims of the geographer, do not necessarily require a descriptively literal account of France's shape. The statement might not aim to assert or describe literally, but do something else entirely. As such,

evaluating it along the lines of truth or falsity will miss something important about the aims of a speaker in uttering it.

To expand on this picture, I turn to explicating Austin's speech act theory.

3. Austin's speech act theory

3.1 Performatives and constatives

Austin first drew our attention to the things we do with words by discussing performative utterances. Austin says of these, "if a person makes an utterance of this sort we should say that he is *doing* something rather than merely *saying* something" (1970, 235). Imagine a speaker utters 'I promise to return my referee report in two weeks' during the peer review process. In making this speech act, Austin claims the speaker does not describe an internal act she has concurrent to her utterance. Instead, in making that utterance, the speaker just is performing the act of promising thereby committing herself to actions related to the timely review of papers.

While promising has no special connection to truth and falsity, it still must meet what Austin calls felicity conditions to be happy or unhappy. In order to promise to return their referee report in two weeks successfully, the speaker must meet the sincerity condition of forming an intention to do so, even if they are not describing "some inward spiritual act of promising" (236). The speaker must also be in a position to follow through on their intention. Thus, there is unhappiness in the speech act if the speaker promises knowing full well other commitments will prevent her from returning the report in two weeks. The speaker must also have the authority to make a promise; unless authorized, an editor cannot promise on behalf of a reviewer. There should also exist a convention for making a promise in peer review contexts. Such conventions might allow the speaker to promise without uttering, 'I promise,' e.g., by accepting a request that reads, 'In accepting this review assignment you commit to returning the referee report within such-and-such a time.'

Austin first contrasts performatives with constatives, e.g., descriptive statements or assertions that aim to state something truly or falsely about the world, but which do not seem to perform an action. However, Austin claims describing or asserting is as much an action as promising, even if the felicity conditions for asserting are more closely connected to truth

or falsity. Consider an editor saying of a reviewer, 'They review quickly, and I expect that they will return their review within two weeks.' In saying this, the editor commits herself to providing evidence for her description of the reviewer as quick, and perhaps justifying her expectation that the reviewer's past behavior provides good evidence for future behavior. As Robert Brandom puts this point, "In asserting a claim one not only authorizes further assertions, but commits oneself to vindicate the original claim, showing that one is entitled to make it" (1983, 641). That is, the utterer must be in a position of authority—here in an epistemic sense—with regards to the claim and be ready to perform further speech acts if so prompted. Other felicity conditions of assertions or descriptions include a sincerity condition: an editor uttering our example sentence should believe what they say. Finally, the context of an assertion also shapes its felicity conditions: an editor should utter the sentence in the appropriate circumstances, e.g., as a response to a worry about the speed of the review process. Should these conditions not be met, the speech act might be unhappy even if true.

3.2 Locution and illocution

Austin develops speech act theory to capture the similarities between performatives and constatives. Speech acts like promising and describing have three dimensions: the locutionary content, which is the conventional sense and reference of the uttered sentence; the illocutionary force, which is the use the utterance is put to; and the perlocutionary effects, which are intended and unintended "effects upon the feelings, thoughts, or actions of the audience, or of the speaker, or of other persons" (1962, 101).

Austin's points about the illocutionary dimension of a speech act most clearly capture how one and the same representation might be put to different uses depending on our goals, and how different uses have different felicity conditions despite sharing locutionary content. Consider the sentence, 'This product contains chemicals known to the state of California to cause cancer.' The locutionary content would just consist in the proposition expressed by the sentence as determined by the conventional sense and reference of the words. This content can be common to different illocutionary acts. Someone uttering the sentence could be describing a product, issuing a warning, or explaining why they do not use this particular product but another. Uttering the sentence with the force of a description, the force of a

warning, and the force of an explanation will have similar felicity conditions related to truth and falsity. Namely, the locutionary content should be true or approximately true for an utterance to count as a good description, a good warning, or a good explanation.

However, a warning might be infelicitous in ways a description might not. For example, warnings might be issued only in the case in which some pre-determined level of significant risk at a certain level of exposure is met. In cases where such levels are not met, issuing a warning might be infelicitous. Consider also that uttering such a sentence with the force of an explanation might be called for only if, e.g., someone is prompted to justify their choice of a product that does not contain cancer-causing chemicals over a more easily available and cheaper product that does contain those chemicals. In these last two cases, nonepistemic reasons related to risk, cost-effectiveness, and so on can enter into the evaluation of the happiness of a warning or explanation.[2]

Austin thinks attending to these points combats a form of abstraction that distorts our thinking about the felicity conditions of descriptive statements. He thinks that when examining statements, "we abstract from the illocutionary…aspects of the speech act, and we concentrate on the locutionary" (1962, 144–5). In so doing, "we use an over-simplified notion of correspondence with the facts—over-simplified because essentially it brings in the illocutionary aspect" (145). Such an approach focuses on "the ideal of what would be right to say in all circumstances, for any purpose, to any audience, &c." (145). But, as Austin claims, questions concerning correspondence with the facts brings with it the illocutionary aspect since truth or falsity does not attach to sentences or locutionary content. Instead, truth or falsity is related to particular things speakers do with sentences. Descriptions might be, strictly speaking, true or false, but not recommendations or explanations. In order to know, then, if evaluating a speech act along the true-false dimension is apt, we need to know the illocutionary force of that act. But to know the illocutionary force of the act requires we attend to context, including the aims of both speaker and audience, time and place of utterance, and conventions governing the specific speech situation. In this way, Austin

[2] Any speech act will also have perlocutionary effects, and we might follow Heather Douglas (2009) and Paul Franco (2017) in focusing on the nonepistemic consequences of making false descriptions, giving bad warnings, or explaining unclearly.

argues context and aims are central to determining the illocutionary force of a speech act, and hence to evaluating its felicity or infelicity.

4. Aims-approaches and speech act theory

4.1 Explaining and understanding

Scientific practice might seem to deal in paradigmatically constative speech acts, e.g., descriptions. Such speech acts are, to varying degrees, evaluable along dimensions of truth or falsity in ways we might question the relevance of speech act theory to philosophy of science. That is, we might say that scientific practice just is a case in which abstracting away from the illocutionary force of an utterance to focus on locutionary content is appropriate. For example, Austin says that "perhaps with mathematical formulas in physics books…we approximate in real life to finding" speech acts where focusing on the locutionary content is appropriate (1962, 145). If scientific practice aims at timeless truths holding across all contexts independent of the sorts of aims and interests of speakers and audiences necessary to evaluating the felicity or infelicity of speech acts, then it seems speech act theory is irrelevant to philosophy of science.

Yet, as Austin points out, "When a constative is confronted with facts, we in fact appraise it in ways involving the employment of a vast array of terms which overlap with those that we use in the appraisal of performatives. In real life, as opposed to the simple situations envisaged in logical theory, one cannot always answer in a simple manner whether it is true or false" (141–2). Consider again 'France is hexagonal.' Austin asks, "How can one answer…whether it is true or false that France is hexagonal? It is just rough, and that is the right and final answer to the question of the relation of 'France is hexagonal' to France. It is a rough description; it is not a true or false one" (142). Though rough, it is still open to evaluation. We can ask if it is in accord with conventions governing estimations and if this estimation serves the purposes and interests of the speaker and their audience at the time of utterance. 'France is hexagonal' can count as felicitous even if rough and not literally true because it aims at something other than truth.

Austin claims that many of our apparently constative speech acts are evaluable along similar dimensions given that they also confront facts in similarly rough ways. McKaughan

makes a related point about scientific speech acts. He argues that certain speech acts central to scientific practice like "conjecturing, hypothesizing, guessing and the like often play a role in scientific discourse that serves neither to assert that an hypothesis is true nor to express such a belief" (2012, 89). Moreover, as mentioned in §2, the picture of scientific practice as concerned solely with the truth is challenged, among other places, in work on explanation, and also in values in science. For example, when looking at the role particular acts or patterns of explaining play in scientific discourse we might focus not on the locutionary content of an explanatory speech act, but on the ways "explanatory discourse…functions to sculpt and subsequently perpetuate communal norms of intelligibility" (Woody 2015, 81). In focusing on this aspect of explaining, we might find, for example, that "the ideal gas law's role in practice is not essentially descriptive, but rather prescriptive; by providing selective attention to, and simplified treatment of, certain gas properties (and their relations) and ignoring other aspects of actual gas phenomena, the ideal gas law effectively instructs chemists in how to think about gases as they are characterized within chemistry" (82). In other words, the ideal gas law, in practice, does not have the force of a descriptive speech act, but lays down a rule of sorts guiding the investigation of gases.[3] The success of acts of explaining from this perspective will have less to do with accurately describing actual gases, but the way they facilitate, say, the education of new scientists or increase understanding of related phenomena, e.g., "by laying foundation for the concept of 'temperature'" beyond "the subjective, inherently comparative quality of human perception" (82). An act of explaining that fails to achieve pedagogical aims or fails to increase understanding of related phenomena might be infelicitous even if the locutionary content of that act confronts the facts in the right way to count as approximately true.

On this point about the ways explanations might increase understanding without describing, Potochnik claims "that what best facilitates understanding is not determined solely by the relationship between a representation and the world" (2015, 74). An idealized explanation like the ideal gas law is not defective because it fails to fully describe all the

[3] About universal generalizations Austin writes, "many have claimed, with much justice, that utterances such as those beginning 'All…' are prescriptive definitions or advice to adopt a rule" (1962, 143). Austin does not fully endorse this suggestion.

possible causal factors at play in the behavior of actual gases. Though literally false, an idealization might be successful insofar as it "secure[s] computational tractability" or successfully isolates "all but the most significant causal influences on a phenomenon" (71). In so doing, we increase our understanding by facilitating "successful mastery, in some sense, of the target of understanding" or "by revealing patterns and enabling insights that would otherwise be inaccessible" (72). Indeed, pointing out all the ways in which the ideal gas law fails to hold for actual gases or is literally false as a description might hinder the use of explanations in scientific discourse to provide "shared exemplars that function as norms of intelligibility" (Woody 2015, 84).

In a related vein, Potochnik argues, "Because understanding is a cognitive state, its achievement depends in part on the characteristics of those who seek to understand," including both the speaker and the audience (2015, 74). In evaluating an act of explaining, we should look at how the speaker's interest has shaped the focus of their explanation and also how the explanation increases an audience's understanding, where this involves considering the audience's interests in seeking an explanation. An explanation that fails to be relevant to the audience or fails to increase their understanding or guide their thinking about related phenomena, but that nonetheless has locutionary content that is approximately true, might count as infelicitous.

4.2 Values and science

On the views of explaining canvassed, the aims of generating literally true descriptions of the world come apart from, say, explaining and understanding the most important causal factors at play for a given phenomenon. Now, as the aims approach to the proper role for nonepistemic values in scientific practice emphasizes, explaining and describing do not exhaust the goals of scientific practice. The aims approach focuses on the ways "scientific decision-making, including methodological choices, selection of data, and choice of theories or models, are...a function of the aims that constitute the research context" (Intemann 2015, 218). Given that the research context includes social, political, and moral considerations, the aims of science can just as well be understood in nonepistemic ways as it can be understood in epistemic ways.

Consider, for example, the American Geophysical Union's position statement on human-induced climate change. At the end of their statement, they claim, "The community of scientists has responsibilities to improve overall understanding of climate change and its impacts. Improvements will come from pursuing the research needed to understand climate change, working with stakeholders to identify relevant information, and conveying understanding clearly and accurately, both to decision makers and to the general public" (American Geophysical Union 2013). Here, I focus on the claim that scientists have responsibilities to improve the understanding of policymakers and the general public, and drawing upon the aforementioned work on explaining, think about how adopting this aim shapes the felicity conditions of explanatory speech acts directed at the audiences mentioned.

Notice that the position statement distinguishes the research necessary to understand climate change from conveying that understanding to policymakers and the general public. The sense in which these different activities come apart from one another and have different success conditions can be made sense of, in part, by focusing on the audience to whom scientists are speaking. We saw that for Potochnik (2016) understanding is a cognitive state that depends on the abilities and interests of those who are explaining and those to whom explanations are directed. In communicating to policymakers and the general public, scientists should consider the interests of the speaker in asking for an explanation as well as their level of knowledge regarding the phenomenon in question, in this case, climate change. In so doing, scientists might find that a description that aims to describe climate change in all its complexity might not serve these aims well. Instead, scientists might aim for an explanation that, though omitting descriptive complexity, draws upon models that represent those causal factors related to the audience's interests in a way that is cognitively accessible and helps guide the public in thinking more generally about climate change.

On this point, the American Geophysical Union's position statement maintains scientists ought to enlist the help of stakeholders in identifying potentially relevant information to their research. This is a point Intemann makes in developing the aims approach. She says of climate science, "[T]he aim is not only to produce accurate beliefs about the atmosphere, but to do so in a way that allows us to generate useful predictions for protecting a variety of social, economic and environmental goods that we care about" (2015,

219). In the view of the American Geophysical Union, in order to do this well, scientists ought to consult with relevant stakeholders and policymakers regarding what they value. Thus, for example, if stakeholders and policymakers communicate worries about extreme weather events and "how to adapt to 'worst case scenarios,' then models able to capture extreme weather events should be preferred" to those models that "anticipate slow gradual changes" (Intemann 2015, 220). Notice that in making such a decision, the grounds for choosing models able to represent aspects of climate change relevant to stakeholders' interests are nonepistemic rather than epistemic, e.g., generating predictions useful for protecting goods the general public cares about. Insofar as the representations or explanations generated do not meet these goals because they are unrelated to stakeholders' interests, the attendant speech acts might very well be infelicitous even if they describe some related phenomenon more or less accurately.

Both points about pitching explanations at a level that is cognitively accessible and choosing models for representing climate change phenomena in ways sensitive to stakeholders' interests illustrate a point Austin makes about the importance of uptake to successfully performing a speech act. Austin claims, "Unless a certain effect is achieved, the illocutionary act will not have been happily, successfully performed….I cannot be said to have warned an audience unless it hears what I say and takes what I say in a certain sense….Generally the effect amounts to bringing about the understanding of the meaning and force of the locution" (1962, 116). In aiming to convey understanding through explaining relevant aspects of climate change to decision makers and the general public, a speaker should consider the interests, background knowledge, and cognitive resources of their audience. Insofar as scientists fail to do so in explaining to the general public, even if the locutionary content that comprises their speech act approximates truth, they will not secure uptake in the sense of generating understanding in their audience. As such, their speech act will be infelicitous.

Of course, a scientist's explaining something to their audience will also be infelicitous if it is based on inaccurate information or extrapolates from what is known to their audience's interests in unjustified ways. However, this does not mean that if scientists aim to convey understanding to the public they should stick solely to descriptive claims. As

Elliott emphasizes in discussing how scientists should best communicate uncertainty to the public, "It does little good to expect scientists to provide unbiased information to the public if their pronouncements are completely misinterpreted or misused by those who receive them" (2017, 89). Similarly, "members of the public might not be able to 'connect the dots'" between scientists' descriptive speech acts and the ways those are relevant to their interests; insofar as scientists do not explain with the aims of conveying understanding—which as Potochnik argues, comes apart from describing the world truly in all its complexity—the public "would be left wondering what [the descriptions] might mean" (88). Thus, if scientists are to meet responsibilities the American Geophysical Union claims they have with regard to conveying understanding to the general public, those scientists should communicate using speech acts best able to secure uptake in the general public. This involves considering the interests and cognitive resources of the general public in ways that shape the felicity conditions of the speech acts beyond truth and falsity.

5. Conclusion

I argued speech act theory can tie together a few threads in recent work on explaining and values in science that share in common a shift in focus from descriptive propositions to things scientists do with words. Some of those things, like explaining, also seem the sorts of speech acts appropriate for fulfilling aims scientists have other than describing the world literally, like conveying understanding to the public and policymakers. Insofar as successfully fulfilling these aims involves explaining, and insofar as acts of explaining that secure uptake require attention to the nonepistemic interests and cognitive resources of speaker and audience, our attention is drawn towards ways explanatory speech acts can be happy or unhappy beyond describing truly or falsely. Future work will aim to delineate these felicity conditions in greater detail with an eye towards revealing further nonepistemic dimensions of evaluation.

References

Achinstein, Peter. 1977. "What is an Explanation?" *American Philosophical Quarterly* 14(1):1–15.

American Geophysical Union. 2013. "Human-Induced Climate Change Requires Urgent Action." https://sciencepolicy.agu.org/files/2013/07/AGU-Climate-Change-Position-Statement_August-2013.pdf

Austin, J.L. 1962. *How to Do Things With Words*. Ed. J.O. Urmson. Oxford: Oxford University Press.

----. 1970. "Performative Utterances." *Philosophical Papers*, 2nd edition. Eds. J.O. Urmson and G.J. Warnock. Oxford: Oxford University Press: 233–252.

Brandom, Robert. 1983. "Asserting", *Nous* 17(4):637–650.

Cartwright, Nancy. 1980. "The Truth Doesn't Explain Much." *American Philosophical Quarterly* 17(2):159–163.

Chang, Hasok. 2014. "Epistemic Activities and Systems of Practice: Units of Analysis in Philosophy of Science After the Practice Turn." *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science*, eds. Léna Soler, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost. New York: Routledge: 67–79.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, Kevin. 2017. *A Tapestry of Values*. New York: Oxford University Press.

Elliott, Kevin C. and Daniel J. McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81(1):1–21

Franco, Paul L. 2017. "Assertion, Nonepistemic Values, and Scientific Practice." *Philosophy of Science* 84(1):160–180.

Intemann, Kristen. "Distinguishing Between Legitimate and Illegitimate Values in Climate Modeling." *European Journal of the Philosophy of Science* 5:217–232.

McKaughan, Daniel J. 2012. "Speech acts, attitudes, and scientific practice: Can Searle handle 'Assuming for the sake of Hypothesis'?" *Pragmatics and Cognition* 20:1:88–106.

Potochnik, Angela. 2015. "The Diverse Aims of Science." *Studies in History and Philosophy of Science Part A* 53:71–80

----. 2016. "Scientific Explanation: Putting Communication First." *Philosophy of Science*, 83:721–732.

Woody, Andrea. 2015. "Re-orienting discussions of scientific explanation: A functional perspective." *Studies in History and Philosophy of Science Part A* 52:79–87.

Universality Reduced

Alexander Franklin*†

October 2018

Forthcoming in *Philosophy of Science: Proceedings of the PSA 2018*

Abstract

The universality of critical phenomena is best explained by appeal to the Renormalisation Group (RG). Batterman and Morrison, among others, have claimed that this explanation is irreducible. I argue that the RG account is reducible, but that the higher-level explanation ought not to be eliminated. I demonstrate that the key assumption on which the explanation relies – the scale invariance of critical systems – can be explained in lower-level terms; however, we should not replace the RG explanation with a bottom-up account, rather we should acknowledge that the explanation appeals to dependencies which may be traced down to lower levels.

1 Introduction

While universality is best explained with reference to the Renormalisation Group (RG), that explanation is nonetheless reducible. The argument in defence of this claim is of philosophical interest for two reasons: first, the RG explanation of universality has been touted by Batterman (2000, 2017) and

*alexander.a.franklin@kcl.ac.uk

†I am grateful to Eleanor Knox, and to the audience of the IMPS 2018 conference in Salzburg for helpful comments. This work was supported by the London Arts and Humanities Partnership.

Morrison (2012, 2014) as a significant impediment to reduction. Second, universality is a paradigm instance of multiple realisability (MR) in the philosophy of physics; as such it is regarded as irreducible by those who accept the multiple realisability argument against reduction. My account charts a middle course: I deny claims that RG explanations are irreducible, and I deny that universality is *best* explained from the bottom up.

The view of reduction advocated here is non-eliminativist; the best explanations are often higher-level explanations: such explanations are more parsimonious, more robust, and have broader applicability than lower-level explanations. In general, such higher-level explanations ought not to be replaced by lower-level explanations, rather the parts of theories on which such explanations rely may be understood in lower-level terms; reducible explanations satisfy the following two conditions: (a) each higher-level explanatory dependency is explained by or derived from a lower-level dependency, and (b) the abstractions involved in constructing the higher-level explanations are justified from the bottom up.[1]

In §2 I outline the RG explanation of universality. Although my reductive claims may generalise, I focus exclusively on the field-theoretic approach to the RG.[2] I claim that this explanation follows a general formula for explaining multiply realised phenomena. §3 considers the arguments of Batterman and Morrison, and analyses their force against any putative reduction.

In §4 I note that the RG explanation is a higher-level explanation. As it is less contentious that the common features of each universality class are reducible, I simply assume that that's the case in this paper. The nub of the debate rests on the RG: I show that the RG arguments rely on the assumption of scale invariance and the abstractions engendered by that assumption. I argue that the applicability of this assumption may be explained from the bottom up. Thus, I claim, that my reduction satisfies (a) and (b) above.

[1] While I expect the claims in this paper to be compatible with many different accounts of explanation, they are most straightforwardly cashed out on an interventionist approach – see Woodward (2003).

[2] See Franklin (2018) and Mainwood (2006) for arguments that only this approach provides an adequate explanation of universality.

2 The RG Explanation of Universality

'Universality' refers to the phenomenon whereby diverse systems exhibit similar scaling behaviour on the approach to a continuous phase transition. Continuous phase transitions occur at the critical temperature, a point beyond which systems no longer undergo first-order phase transitions.[3] The approach to this phase transition can be very well described by power laws of the form $a_i(t) \propto t^{\alpha}$ where t is proportional to the temperature deviation from the critical temperature and α is the critical exponent – a fixed number which leads to a characteristic curve on temperature-density plots.[4]

Different physical systems can be categorised into universality classes: members of the same class have identical critical behaviour – the same set of critical exponents $\{\alpha, \beta, ...\}$ for several power laws – while their behaviour away from the critical point and microscopic organisation may be radically different. For example, fluids and magnets are in the same universality class despite otherwise having totally different chemical and physical properties.

Each physical system which exhibits critical phenomena may be described at the critical point by the same mathematical object – the Landau-Ginzburg-Wilson (LGW) Hamiltonian. That Hamiltonian will include the features – the symmetry and dimensionality – which sort these systems into their universality classes. The RG argument demonstrates that the LGW Hamiltonian applies to a wide range of systems at the critical point by showing that any additional operators which may be appended to that Hamiltonian will fall away on approach to criticality, where only the central LGW operators will remain. The following steps are essential to the explanation thus on offer:[5]

1. Define the effective Hamiltonian for your system of interest:
   - (i) Specify the order parameter with symmetry and dimensionality.
   - (ii) Specify the central operators of the LGW Hamiltonian.

[3] Note that not all continuous phase transitions are associated with first-order phase transitions in this way.

[4] E.g. the specific heat (in zero magnetic field) c scales as $c \sim (t^{-\alpha})/\alpha$ as $t \to 0$ where $t = \frac{T - T_c}{T_c}$.

[5] To see a full account of the physics of universality and details of the RG see Binney et al. (1992) and Fisher (1998); the philosophical aspects of such an explanation are discussed in detail in Batterman (2016) and Franklin (2018).

(iii) Specify operators in addition to the terms in the LGW Hamiltonian.

2. Apply the RG transformations to that Hamiltonian.
3. Examine the flow towards fixed points in the critical region and note that some operators are irrelevant to the critical behaviour.
4. Thus divide the set of operators into subsets: 'relevant', 'irrelevant' and 'marginally relevant'.
5. Repeat for other systems of interest.

In order to explain universality we must identify commonalities between the different systems in the same universality class – 1(i) and 1(ii) above – and show that such commonalities are sufficient for the common behaviour – 2-4 above. Although 1(iii) can't, in general, be done explicitly, the explanation only depends on the RG demonstration that all distinguishing features are irrelevant – it's not necessary to say exactly which those distinguishing features are. As discussed below, the infinities which are central to some of the anti-reductionist arguments feature in steps 3 and 4.

Overall the explanation takes the following form: consider a universality class composed of four different physical systems A-D. Each of A-D is described in step 1 by an effective Hamiltonian; effective Hamiltonians are ascribed to systems on the basis of various theoretical and empirical data. The RG explanation of universality, by virtue of steps 2-4, tells us that all the details which distinguish A-D, i.e. their irrelevant operators, are, in fact, irrelevant to the critical phenomena. Thus we have an explanation for how otherwise different systems exhibit the same phenomena at the critical point. This explanation relies, of course, on the RG transformations which allow for the categorisation of certain operators as irrelevant.

Importantly, this explanation takes the form of a general explanation of multiply realised phenomena: such phenomena are explained if commonalities are identified among the realisers and these are shown to be sufficient for the multiply realised phenomena to occur. Note that such explanations may be higher level and nothing written so far establishes their reducibility.

3 Anti-reductionist Arguments

Batterman (2000, 2017) and Morrison (2012, 2014) offer two arguments in defence of the view that the explanation just outlined is irreducible. The more general argument is that universality, *qua* instance of multiple realisability, is irreducible because multiple realisability requires abstracted explanations of a particular form.

However, one goal of this paper is to demonstrate that just such abstracted explanations may be reducible. Insofar as my reduction of the RG explanation goes through, we are thus faced with a dilemma: either some instances of MR are, in principle, reducible, or universality is not a case of MR. While I would opt for the former horn, nothing in the rest of the paper hangs on that choice.

The second anti-reductionist argument is much more specific to the case at hand and involves various demonstrations that the RG explanation requires infinities which are inexplicable from the bottom up. As noted by Palacios (2017), two different limits are invoked in the case of continuous phase transitions – the thermodynamic limit and the limit of scale invariance. There is an extensive literature on the thermodynamic limit as it appears in first order phase transitions; as I see no salient differences between appeal to this limit in the two contexts, I do not discuss this further here – see e.g. Butterfield and Bouatta (2012) for a reductionist account of that limit.[6]

The second limit is discussed by Butterfield and Bouatta (2012), Callender and Menon (2013), Palacios (2017), and Saatsi and Reutlinger (2018), among others, and these papers undermine claims that continuous phase transitions are irreducible. However, they pay insufficient attention to the specific role played by the RG (and by the limit of scale invariance) in establishing the irrelevance of certain details, and it is this role which is crucial to the anti-reductionist arguments.[7]

For Batterman, the RG is required because it allows us to answer the following question:

[6]The reductionist claims made here are conditional on a successful resolution of such issues.

[7]For example, Saatsi and Reutlinger (2018, p. 473) do not consider a counterfactual of the form 'if a physical system S did not exhibit effective scale invariance at criticality, then S would not exhibit the critical phenomena of any universality class' in their list of counterfactuals which the RG account is supposed to underwrite.

> **MR**: How can systems that are heterogeneous at some (typically) micro-scale exhibit the same pattern of behavior at the macro-scale? ...
>
> if one thinks (**MR**) is a legitimate scientific question, one needs to consider different explanatory strategies. The renormalization group and the theory of homogenization are just such strategies. They are inherently multi-scale. They are not bottom-up derivational explanations.
>
> [Batterman (2017, pp. 4, 14-15)]

As further elaborated below, the RG seems to Batterman to preclude "bottom-up derivational explanation" because it requires the following infinitary assumption:

> This [fixed point] is a point in the parameter space which, under τ [the RG transformation], is its own trajectory. That is, it represents a state of a system which is invariant under the renormalization group transformation. Of necessity, such a fixed point has an *infinite correlation length* and so lies on the critical surface S_∞. The singularity/divergence of the correlation length ξ is *necessary*.
>
> [Batterman (2011, p. 1045), original emphasis]

I accept that the RG formalism makes use of infinite limits. The salient question, to borrow Norton's (2012) distinction, is whether such infinities are approximations which allow one to use the more tractable infinitary mathematics to approximate features of the finite systems, or, alternatively, idealisations which describe a distinct infinite system. Claiming that the infinities are idealisations would preclude reduction because the macroscopic system with infinite properties has features which may not be reductively explained.

As Batterman demonstrates, the RG argument rests on the assumption of the infinite correlation length which generates absolute scale invariance. In §4 I claim that the physical systems under consideration are not absolutely scale invariant: in fact, one may abstract from the details of the underlying system insofar as such systems are effectively scale invariant; thus the infinitary assumption is best viewed as an approximation.

While Morrison (2014, p. 1155) likewise focusses on explanations of MR phenomena, she claims that RG explanations are irreducible for a different, but related, reason: the "RG functions not only as a calculational tool but as the source of physical information as well". Morrison (2012) makes a similar argument in relation to symmetry breaking in the physics of superconductors. She argues that, in both cases, top-down constraints play an essential role in the physical descriptions which thus rules out reduction. In the present context, Morrison's views may be understood as taking the RG invocation of scale symmetry to be a necessary physical assumption which cannot be understood from the bottom up. Below I argue that the effective scale invariance on which the RG rests is, in fact, reductively explicable. As such, no top-down organising principles are required and Morrison's claims are deflated.

4 Reducing the RG Explanation

Arguments for the reducibility of the explanation of universality have primarily been targeted at Batterman's claims that infinities are essential to the models used to describe continuous phase transitions. I do not have space to consider these arguments in any detail. Suffice it to say that, in my view, none succeeds in reducing the principal feature of the renormalisation group – the assumption of scale invariance. Thus I focus on that aspect of the RG, and claim that it, too, is reducible.

Furthermore, with the notable exception of Saatsi and Reutlinger (2018), not much attention has been paid to the explanation of universality *per se*. This, of course, makes a difference for MR-based objections to reduction, which raise doubts that a reductionist account could explain why the same phenomenon is exhibited in multiple different systems.

As far as the physics is currently developed, the RG plays an ineliminable role in the explanation of universality: it is the only mathematical framework available to predict the precise extent of observed universality of critical phenomena. If its application were truly mysterious, if we had no idea why it worked, then, infinity or no infinity, this would provide exactly the right kind of failure of explanation on which the anti-reductionist could hang their arguments.

I argue in the following that the applicability of the RG to systems un-

dergoing continuous phase transitions is not mysterious. The RG exploits effective scale invariance to set up equations which tell us how certain properties vary with respect to the variation of other properties. It is a piece of mathematics whose applicability is deeply physical – where the assumptions invoked in applying the RG do not hold, the RG's predictions go wrong.

In order fully to reduce the RG explanation, one also must consider the common features shared by each member of the same universality class, and argue that these, too, are reducible to aspects of the microphysical description. Such arguments have been given by the reductionists mentioned above. The innovation of this paper lies in reducing the RG framework, and the assumptions on which it relies; thus, given space constraints, I do not consider the reduction of the symmetry, dimensionality and representation by common Hamiltonians.

4.1 Reducing the Renormalisation Group

The RG argument rests on the assumption of scale invariance, and this is crucial to the demonstration that a class of operators are irrelevant at criticality. I claim that we can provide a bottom-up explanation of this scale invariance and that, as such, the RG arguments provide a mathematical apparatus for relating scale invariance to the irrelevance of certain details. One can see, heuristically, how scale invariance relates to universality: if the system at criticality is effectively scale invariant then many of that systems' features – those which are scale dependent – will turn out to be irrelevant at criticality, and all that will remain are those shared features such as the symmetry and dimensionality.

To argue that the RG explanation is reducible, I first give a more general characterisation of an RG flow. The calculation of each system's dynamics involves integration over a range of scales and energies. The highest energy (smallest scale) cutoff (denoted Λ) corresponds to the impossibility of fluctuations on a scale smaller than the distance between the particles in the physical system. The RG transformation involves decreasing the cutoff thereby increasing the minimum scale of fluctuations considered. Iterating this transformation generates a flow through parameter space designed to maintain the Hamiltonian form and qualitative properties of the system in question.

The RG transformation $\mathcal{R}$ transforms a set of (coupling) parameters $\{K\}$ to another set $\{K'\}$ such that $\mathcal{R}\{K\} = \{K'\}$. $\{K^*\}$ is the set of parameters which corresponds to a fixed point, defined such that $\mathcal{R}\{K^*\} = \{K^*\}$. This fixed point corresponds to the critical point defined physically. At the fixed point, the RG transformation (which changes the scale of fluctuations) makes no difference. Thus the fixed point encodes the property of scale invariance.

Given the Hamiltonian of one of our models, one can define an RG transformation which generates a flow that allows one to: (i) classify certain of the coupling parameters of the system in question as (ir)relevant to its behaviour near the fixed point, (ii) extract the critical exponents from the scaling behaviour near the fixed point.

The RG may be understood as a mathematical framework for exploring how certain properties vary with changing energy, length-scale, or, by proxy, temperature, on approach to the scale invariant critical point. Philosophical discussions of the RG are occasionally prone to mysterianism, but the RG should be considered to be no different from, for example, the calculus. As Wilson (1975, p. 674) notes: "the renormalization group ... is the tool that one uses to study the statistical continuum limit [the point of scale invariance] in the same way that the derivative is the basic procedure for studying the ordinary continuum limit".

The Hamiltonian which represents the system at the critical point, from which the critical exponents are extracted, is scale invariant at the fixed point – all the scale dependent contributions have gone to zero. Such Hamiltonians are known as 'renormalisable'. As such, the explanation provided below for the effective scale invariance of physical systems at criticality underlies the fact that such systems are well-described by renormalisable Hamiltonians at fixed points.

My argument has two steps: I demonstrate that scale invariance is implicit in the power law behaviour which is intrinsic to universality; then I provide a bottom-up explanation of the effective scale invariance for liquid-gas systems, a story somewhat motivated by the observation of critical opalescence. Thus, I show how scale invariance features in the mathematics – the Hamiltonian's renormalisability and the power laws, and how it features in the observed physics – the critical opalescence is a direct consequence of the bottom-up story.

The universality of critical phenomena lies in the sharing of power laws,

and hence critical exponents, between members of the same universality class. In what sense are such power laws scale-free? As Binney et al. (1992, p. 20) explain, a phenomenon obeying a power law is independent of scale because one could multiply its characteristic scale length by some factor and the ratio of values will remain constant. For example, consider the power law $f_1 = (r/r_0)^\eta$, and its measurement in the range $(0.5r_0, 2r_0)$. The ratio of largest to smallest value will be identical for measurements centred on $r_0, 10r_0, 100r_0$ – it will always be $4^{|\eta|}$, thus one may superimpose all the power laws by a simple change of scale. By contrast, for $f_2 = \exp(r/r_0)$ the ratio of values will change on scale changes.

Such systems are therefore described as scale-free; the RG is used to predict that at the point of scale invariance the heterogeneous features will be irrelevant. So, in order to work out when this framework is applicable, and why it works, we ought to look at each individual system, (for our purposes let's reserve inquiry to liquid-gas and ferromagnetic-paramagnetic systems) and identify the underlying processes which lead to effective scale invariance at the critical point. The following two caveats apply to this proposal for reduction:

First, it might be objected that universality may only be explained if the same processes are identified across all the systems exhibiting the universal behaviour; if that were so, the strategy employed here would be inadequate. However, universality may be explained by demonstrating that two conditions are fulfilled: that all the systems share common features, and that their heterogeneous details are irrelevant. While it's essential that the common features are shared by all the systems, the mechanism by which the heterogeneities are irrelevant may differ, so long as all the heterogeneities in fact end up as irrelevant.

Second, although the power laws and renormalisable Hamiltonians at the fixed point are absolutely scale invariant, the physical systems will, at best, be effectively scale invariant – that is, scale invariant within a certain range of length-scales. That should be acceptable because we know that scale invariance is never exactly true of a system: any real system will be finite and thus violate the assumption at some scale. Moreover, this will not generate empirical problems because the power laws are observed for systems approaching criticality – they are predictions about $T \to T_c$, not $T = T_c$. Thus one should only assume that critical exponents asymptotically approach those predicted at the fixed point. While infinite assumptions are required in order to impose the full scale invariance for RG analy-

sis, I claim that we can explain effective scale invariance for finite systems, and that absolute scale invariance is an approximation invoked to make the mathematics tractable.

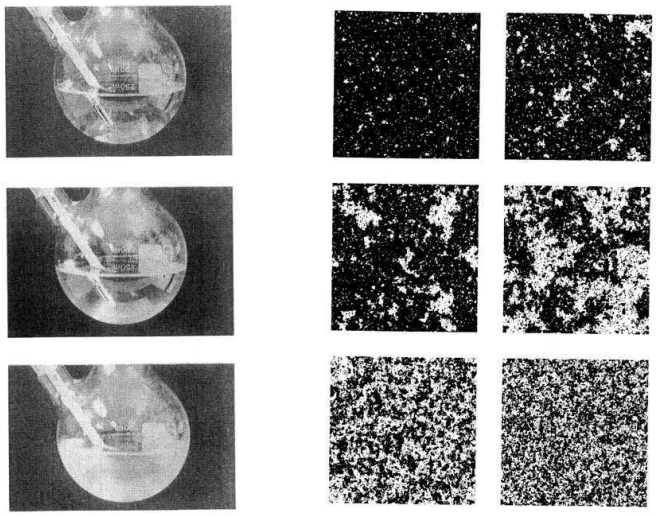
Scale invariance, as it manifests in systems at criticality, is known as 'self-similarity': as scales change the system resembles itself. How do we account for such self-similarity? The critical point, at which a continuous phase transition occurs, corresponds (for liquid-gas systems) to the highest temperature and pressure at which liquid and gas phases can be distinguished.

As is well known, there is a plateau in pressure-volume diagrams, which corresponds to the latent heat (or enthalpy) of vapourisation. This, roughly, is the extra energy needed to break the intermolecular bonds which distinguish liquids from gases and vapours. At the critical point this plateau, and the latent heat of vapourisation vanishes. Now it's difficult precisely to work out the binding energies of the intermolecular bonds. The values for this will be material dependent, and surface tension dependent, and will change at different pressures. But the heuristic argument tells us that the reason the plateau vanishes is because the system has enough temperature, and thus the molecules have sufficient energy to equal the binding energy. The point at which binding energy is exactly matched by kinetic energy will be the critical point.

The isothermal compressibility (κ) is defined as $\kappa_T = \frac{-1}{V}\left(\frac{\partial V}{\partial p}\right)_T$. This corresponds to how much the volume will change (∂V) with a given pressure change (∂p) at fixed temperature (T). As supercritical fluids have far higher compressibility than liquids, and both are present at the critical point, the compressibility diverges. Given, in addition, that the latent heat is zero at criticality, there's nothing to prevent a given bubble expanding arbitrarily. Thus we ought to expect the system to have bubbles of all sizes: this is what is meant by the claim that the system is dominated by fluctuations and has no characteristic scale.[8]

Negligible energy cost for transitions and infinite compressibility leads to self-similarity, and, in certain fluids, the bubbles at all scales lead to a high refraction of visible light. Thus otherwise transparent fluid may become opaque and milky-white. This is known as 'critical opalescence' – see figure 1(a) – and is a visible correlate of a system at criticality.

[8]Note that, for first order phase transitions, the compressibility also diverges; this doesn't lead to scale invariance because latent heat is finite.



(a) (b)

Figure 1: From Binney et al. (1992, pp. 10,19). (a) Critical opalescence is visible when arbitrarily large bubbles form in liquid at criticality. (b) Increasing loss of characteristic scale as $T \to T_c$ in simulations of the Ising model.

Such self-similarity is conceptually crucial to the applicability of the renormalisation group: in order to extract critical exponents from RG equations one identifies a renormalisable Hamiltonian which is scale invariant at the fixed point. Without fluctuations across all scales, systems would fail to be well modelled by such Hamiltonians. The physical argument for diverging fluctuation size justifies the use of a scale invariant mathematical model to represent such systems. Thus, for critical phenomena, the applicability of the RG depends on scale invariance, where this assumption is explicable from the bottom up.

Demonstrating these claims quantitatively is difficult, but the heuristic argument is convincing. Kathmann (2006) reviews theories of the nucleation of gas bubbles in water which generate accurate predictions concerning the rate of bubble growth and the threshold for stability over a range of temperatures; although these models do not reach the critical point, progress is being made.[9]

[9]Constructing exact models is especially difficult because of the fluctuations at a wide range of length scales – precisely the reason that the RG is employed.

Of course, further work could be done to develop these arguments and make them more precise. But there seems to be, in the above, a sound qualitative argument and no in-principle barriers to full derivation. This 'in-principle' ought not to be problematic: we know the relevant physical principles, even if quantitative models are still unavailable.

Moreover, as discussed below, and depicted in figure 1(b), the Ising model allows us quantitatively to predict analogues of the results for liquid-gas systems. While well short of a full explanation, the following discussion illustrates how self-similarity may be reduced for magnetic systems. By treating the Ising model as a stand-in for such systems, a similar kind of reasoning to that given above will go through.

Below the critical point, energy fluctuations will lead to random isolated spin flips. Such flips will be energetically costly and tend to be reversed. The higher the energy, the more likely these are to occur, and if sufficiently many occur then a patch will form, and other spins will have some tendency to align themselves with this patch. However, below the critical point, such patches beyond a certain size will be too costly and spins will overall remain aligned (there is some small probability of net magnetisation flipping, but this is increasingly unlikely further below the critical point).

At the critical point, the energy of the atoms in the lattice is greater than the energetic cost of violating spin alignment, and patches can become arbitrarily large. This results from the latent heat's vanishing and the divergence of the magnetic susceptibility (χ) on approach to the critical point. $\chi_T = \left(\frac{\partial m}{\partial B}\right)_T$ where m is the magnetisation and B represents an external magnetic field. Universality is manifested by the fact that the susceptibility and the compressibility both diverge according to identical power laws with the same critical exponent γ: χ_T, $\kappa_T \sim (T - T_c)^{-\gamma}$. Thus, we have self-similarity and effective scale invariance with bubbles or patches arbitrarily large up to the size of the system.

My aim is to establish the reducibility of the RG relevance and irrelevance arguments. I have demonstrated that the RG is a mathematical procedure that extracts information based on the empirically and theoretically justified assumption of effective scale invariance; this has been shown to be a property shared by different systems at criticality. The key ingredients for effective scale invariance are features of the interactions of neighbouring sub-systems, and the particulate constitution of the materials. While that suggests that these materials are not so different after all, it's worth empha-

sising that the systems which exhibit universal behaviour are nonetheless dissimilar away from the critical point – it's clear that magnets and liquids have many distinct chemical and physical properties.

The assumption of scale invariance plays a crucial role for the RG – it licences the discarding of scale dependent details; it is precisely this discarding of details which ensures that all systems are commonly described at the critical point. Moreover, discarding such details is what gives the higher-level explanation its stability and parsimony. It is thus incumbent on the reductionist to explain how the higher-level RG account is successful despite its leaving out such details. So, the reductionist should identify physical processes at the lower level which ensure the irrelevance of the discarded details.

As argued above, the physical processes in question are exactly those which lead to effective scale invariance. The fluctuations at all scales make it such that the scale-dependent properties which distinguish systems away from criticality are irrelevant at criticality, when the system is effectively scale invariant. We have identified, at the molecular level, the physical mechanisms which prevent variations in the discarded details from leading to changes in the higher-level description of the system. As such, we are assured that the explanatory value of the higher-level explanation is a consequence of features of the lower-level system.

One upshot of this reductionist account is that we may specify the conditions under which the higher-level description remains a good one. The discarded details are irrelevant while the large scale fluctuations – the bubbles or patches – dominate the physics. As we move to systems which are less scale invariant, as the bubbles die down, the critical point becomes a less accurate description and each system in the class will start to exhibit distinct behaviour. This is reflected in the fact that the macroscale RG description only derives the shared behaviour at the fixed point of scale invariance and predicts distinct behaviour away from the fixed point.

I end this section with the following intuitive physical gloss on the RG explanation: "[b]ecause the fluctuations extend over regions containing very many particles, the details of the particle interactions are irrelevant, and a great deal of similarity is found in the critical behavior of diverse systems" (A. L. Sengers, Hocken, and J. V. Sengers (1977, p.42)). Since we can explain the wide-ranging fluctuations from the bottom-up, the RG explanation of universality is reducible.

5 Conclusion

The field-theoretic RG framework, together with the common features of physical systems in the same universality class, explains how those systems all display the same critical phenomena when undergoing continuous phase transitions. That explanation is a higher-level explanation.

That higher-level RG explanation is nonetheless reducible. That is, we may explain in terms of the microstructure of each system how it is that each aspect of the higher-level explanation is explanatory. We may, in particular, show why the RG categorisation of operators as relevant and irrelevant works. That division depends on the assumption of scale invariance, and the assumption of scale invariance is justifiable when systems are effectively scale invariant at criticality.

The anti-reductionist claim that universality is MR, and MR is essentially irreducible has been undermined by demonstrating that we may arrive at a bottom-up understanding of the common features and of what makes such features sufficient for the common behaviour.

The further argument that the use of the infinite limit imposes an irreducible divide between the higher-level and lower-level models has similarly been countered: while we move to the infinite limit in order to make the mathematics simpler, the effective scale invariance can be shown to follow from details of the particle interactions at criticality – that's what identifies the critical point and allows us to make the corresponding abstractions from scale dependent details. Provided with this bottom-up explanation, there is no further reason to claim that the infinite limit is an idealisation rather than an approximation: for we have explained from the bottom up how the system is approximately self-similar.

One upshot of this discussion is that the RG is not to be regarded as mysterious, or, somehow, as the source of physical information. It is applicable only insofar as the systems to which it is applied have the relevant properties, and their having such properties may be reductively explained.

References

Batterman, Robert W. (2000). "Multiple Realizability and Universality". In: *The British Journal for the Philosophy of Science* 51.1, pp. 115–145.

Batterman, Robert W. (2011). "Emergence, singularities, and symmetry breaking". In: *Foundations of Physics* 41, pp. 1031–1050. DOI: 10.1007/s10701-010-9493-4.

— (2016). "Philosophical Implications of Kadanoff's work on the Renormalization Group". In: *Journal of Statistical Physics (Forthcoming)*.

— (2017). "Autonomy of Theories: An Explanatory Problem". In: *Noûs*. DOI: 10.1111/nous.12191.

Binney, James J. et al. (1992). *The Theory of Critical Phenomena: an Introduction to the Renormalization Group*. Clarendon Press, Oxford.

Butterfield, Jeremy and Nazim Bouatta (2012). "Emergence and Reduction Combined in Phase Transitions". In: *AIP Conference Proceedings* 1446, pp. 383–403. DOI: 10.1063/1.4728007.

Callender, Craig and Tarun Menon (2013). "Turn and Face the Strange ... Ch-ch-changes Philosophical Questions Raised by Phase Transitions". In: *The Oxford Handbook of Philosophy of Physics*. Ed. by Robert W. Batterman. Oxford University Press, pp. 189–223.

Fisher, Michael E. (1998). "Renormalization group theory: Its basis and formulation in statistical physics". In: *Reviews of Modern Physics* 70.2, p. 653.

Franklin, Alexander (2018). "On the Renormalization Group Explanation of Universality". In: *Philosophy of Science* 85.2. DOI: 10.1086/696812.

Kathmann, Shawn M. (2006). "Understanding the chemical physics of nucleation". In: *Theoretical Chemistry Accounts* 116.1, pp. 169–182. DOI: 10.1007/s00214-005-0018-8.

Mainwood, Paul (2006). "Is More Different? Emergent Properties in Physics". PhD thesis. University of Oxford.

Morrison, Margaret (2012). "Emergent Physics and Micro-Ontology". In: *Philosophy of Science* 79.1, pp. 141–166. DOI: 10.1086/663240.

— (2014). "Complex Systems and Renormalization Group Explanations". In: *Philosophy of Science* 81.5, pp. 1144–1156. DOI: 10.1086/677904.

Norton, John D. (2012). "Approximation and Idealization: Why the Difference Matters". In: *Philosophy of Science* 79.2, pp. 207–232.

Palacios, Patricia (2017). *Phase Transitions: A Challenge for Reductionism?* URL: philsci-archive.pitt.edu/13522/.

Saatsi, Juha and Alexander Reutlinger (2018). "Taking Reductionism to the Limit: How to Rebut the Antireductionist Argument from Infinite Limits". In: *Philosophy of Science* 85.3, pp. 455–482. DOI: 10.1086/697735.

Sengers, Anneke Levelt, Robert Hocken, and Jan V. Sengers (1977). "Critical-point universality and fluids". In: *Physics Today* 30.12, pp. 42–51.

Sober, Elliott (1999). "The multiple realizability argument against reductionism". In: *Philosophy of Science*, pp. 542–564.

Wilson, Kenneth G. (1975). "The renormalization group: Critical phenomena and the Kondo problem". In: *Reviews of Modern Physics* 47.4, pp. 773–840.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. Oxford University Press.

Title: There Are No Ahistorical Theories of Function

Author: Justin Garson

Abstract: Theories of function are conventionally divided up into historical and ahistorical ones. Proponents of ahistorical theories often cite the *ahistoricity* of their accounts as a major virtue. Here, I argue that none of the mainstream "ahistorical" accounts are actually ahistorical. All of them embed, implicitly or explicitly, an appeal to history. In Boorse's goal-contribution account, history is latent in the idea of statistical-typicality. In the propensity theory, history is implicit in the idea of a species' natural habitat. In the causal role theory, history is required for making sense of dysfunction. I elaborate some consequences for the functions debate.

Keywords: Philosophy of biology; biological function; selected effects; causal role; fitness contribution

Address: Department of Philosophy, Hunter College of the City University of New York, 695 Park Ave., New York, NY 10065

Email: jgarson@hunter.cuny.edu

1. Introduction

Theories of function are conventionally divided up into two main categories, historical and ahistorical (or backwards-looking and forwards-looking). The selected effects theory (Neander 1983, 1991; Millikan 1984) is an example of a *historical* theory, but there are other historical theories, including some versions of the organizational theory (McLaughlin 2001), and the weak etiological theory (Buller 1998). *Ahistorical* theories include Boorse's goal-contribution account (1976; 1977; 2002), the propensity theory (Bigelow and Pargetter 1987), and the causal role theory (Cummins 1975; Hardcastle 2002; Craver 2001; 2013). In the 1970s and 1980s, it was common to see these two sorts of theories as competing with each other, though more recently, philosophers of biology have generally adopted a pluralistic stance, and see them as capturing different aspects of real biological usage (OMITTED). Still, the validity of the basic distinction has never been seriously challenged.

Many proponents of ahistorical theories have argued that we should accept their theories precisely *on account of* their being ahistorical. In other words, their alleged ahistoriticity is often held up as a significant virtue of their theories, and a strong reason to prefer them to historical theories (or at least a strong reason to think they capture a significant strand of ordinary biological usage). There are two arguments along these lines. The first argument appeals to bald intuition, and says that it's just obvious that functions don't always need history. One fanciful variant of this argument appeals to science fiction cases, like swamp creatures, instant lions, and randomly-generated worlds (e.g, Boorse 1976, 74; Bigelow and Pargetter 1987, 188). But one doesn't have to go as far as science fiction to find plausible cases of ahistorical functions in biology. Many philosophers have a strong intuition that, the very first time a new biological trait emerges and begins to benefit the organism, it has a *function* even if it was never selected for (e.g., Boorse 2002, 66; Bigelow and Pargetter 1987, 195; Walsh and Ariew 1996, 498). The second argument, which is closely related, appeals to ordinary biological usage, not intuition. It says that historical theories run against the way biologists ordinarily think and talk about functions. At least sometimes, when biologists attribute functions to traits, they do not *cite* or *refer to* or *think about* history or evolution (e.g., Godfrey-Smith 1993, 200; Amundson and Lauder 1994, 451; Walsh 1996, 558; Boorse 2002, 73). Hence, ahistorical theories capture important strands of real biology.

In light of the above, my thesis might come as a bit of a shock. I claim that *there are no ahistorical theories of function* – or, to put it more precisely, the mainstream versions of the allegedly ahistorical theories on the market are not actually ahistorical. If we poke and prod at those theories a bit, a historical element falls out, like contraband stashed away in a suitcase. In Boorse's version of the goal-contribution account, history is explicitly embedded in his notion of a *statistically-typical* contribution to fitness. In the propensity account, history is embedded, a little less explicitly, in the idea of a species' *natural habitat*. Finally, I claim that the only way the causal-role theorist can hope to make sense of dysfunction is to appeal to history.

If this thesis is correct – that there are no ahistorical theories of function – three consequences immediately follow. First, we need to jettison this whole way of dividing up theories of function. The distinction between etiological and non-etiological theories serves us much better, as I'll describe in the conclusion. The distinction between etiological and non-etiological theories doesn't map onto the distinction between historical and ahistorical theories; rather, *these are two ways of being historical*. Second, given that there are no ahistorical views, a good portion of the arguments that have been put forward to date for these theories (those I mentioned above) are unsound. A third consequence is that one popular way of thinking about function pluralism must fail. This sort of pluralist wishes to sort all biological usage under two main umbrella theories, the selected effects theory and the causal role theory. An argument for this sort of pluralism is that it mirrors the two main uses of "function" in biology, the historical sense and the ahistorical sense. If I'm right, this incarnation of the pluralist project can't possibly work.

Before I move on, there is one big qualification I must get out of the way. One could, just for fun, *invent* a purely ahistorical theory of function. One could assert, for example, that *all* of a trait's effects are its functions. This theory (pan-functionalism?) would be ahistorical, to be sure, since even if the world were created two seconds ago in pretty much its present form, things would still have effects, and so they'd still have functions. In fact, sometimes scientists actually *do* use the word "function" synonymously with "effect." They say things like, "climate change is a *function* of deforestation," or "poor academic performance is a *function* of malnutrition." Clearly, there are some ahistorical uses of "function." But this isn't the ordinary biological use, which the theories I cite above are trying to capture.

So, I need to amend my thesis slightly. Instead of saying that there are no ahistorical theories of function, I want to say that any theory of function that satisfies two very minimal, very traditional, and largely uncontroversial, adequacy conditions, is *also* a historical theory. First, the theory should capture some distinction between functions and accidents (the function of the nose is to help us breathe but not hold up glasses). Second, the theory should capture the possibility of malfunctioning or dysfunction. If my heart seizes up due to cardiac arrest, it's failing to perform its function or it's dysfunctional. All of the theorists I engage with in this paper purport to satisfy these two adequacy criteria, or something like them, so I'm not begging any questions by insisting on these conditions.

Here's the plan for the rest of the paper. There are five sections. After the introduction, I'll turn to Boorse's version of the goal-contribution theory, and show how it explicitly contains a historical element (Section 2). Then I'll turn to the propensity theory and show how it contains a reference to history, buried inside the idea of a trait's *natural habitat* (Section 3). I will then show how the causal-role theory, if it is to make any sense of dysfunction, must include a reference to history (Section 4). In the conclusion (Section 5), I'll reiterate the big consequences for thinking about functions and suggest a better way of dividing up theories of function.

2. Boorse's Goal-Contribution Account

Boorse's view (1976; 1977; 2002), at the most general level, is a goal-contribution account. It holds that a trait's function is just its contribution to a goal. The plausibility of this view stems from its ability to reconcile artifact and biological functions in a single theory: the function of an artifact depends on its contribution to the goal of its user; the function of a biological trait depends on its contribution to the goal of the organism or the lineage. Here, I'll focus on the subclass of functions he calls *physiological* functions.

For Boorse, the *physiological* function of a trait is its species-typical contribution to the survival and reproductive prospects of an organism (1977, 555; 2002, 72). (To be more precise, Boorse carves up species into subgroups based on age and sex; the function of a trait is its typical contribution to fitness within the members of that subgroup.) Though he doesn't define a corresponding notion of *dysfunction*, he defines a closely related notion of *disease*: a disease is simply a state that "reduces one or more functional abilities below typical efficacy."

One of Boorse's arguments for the superiority of his theory over Wright's (1973) etiological approach, and the selected effects theory of Millikan (1984) and Neander (1983), is that his approach *makes no reference to history*. He advances two arguments for the value of this ahistorical approach; one appeals to ordinary biological usage, and the other appeals to intuition. First, he says, the goal-contribution account fits ordinary biological usage: "in talking of physiological functions, they [that is, pre-Darwinian biologists] did not mean to be making historical claims at all. They were simply describing the organization of a species as they found it" (1976, 74). The same is true of current physiologists, who have "*no thought* of explaining [a trait's] history" when they assign functions to them (Boorse 2002, 73, emphasis mine). All historical theories of function simply miss how physiologists have always used the word "function." His second argument appeals to intuition. He says that intuition revolts against putting history into functions, as attested to by his instant lions case. If the lion species sprang into existence by "unparalleled saltation," one would *not* say that the parts of lions don't have functions (ibid.; also see Boorse 2002, 75). Again, functions can't be historical.

Neander (1991, 182) raised a now-famous objection against Boorse; she pointed out that Boorse's view, as it stands, can't make sense of pandemic disease: "dysfunction can become widespread within a population…A statistical definition of biological norms implies that when a trait standardly fails to perform its function, its function ceases to be its function; so that if enough of us are stricken with disease (roughly, are dysfunctional) we cease to be diseased, which is nonsense." Pandemic diseases, moreover, don't just occupy the realm of science fiction, as in P. D. James' *The Children of Men*. UV radiation poisoning in anurans is a good example of pandemic dysfunction. Sadly, climate change might create many more pandemic dysfunctions very soon. A good theory of function should at least allow for the *conceptual* possibility that all, or most, tokens of a certain trait in a certain species are dysfunctional (or as Boorse prefers, "diseased").

Intriguingly, Boorse doesn't deny the possibility of pandemic disease. Instead, he says that in order to make sense of pandemic disease, one has to appreciate function's

historical depth. Specifically, he says that when we consider what is "statistically typical" for a trait, we cannot just look at what is typical right now. Rather, we have to consider what is typical within a long slice of time that extends far back into the past: "Obviously, some of the species' history must be included in what is species-typical. If the whole earth went dark for two days and most human beings could not see anything, it would be absurd to say that vision ceased to be a normal function of the human eye (2002, 99)." He tells us that this time-slice should be longer than "a lifetime or two," and might include "millennia."

This is an extraordinary admission, given that much of Boorse's core argument *for* his view was propped up on the claim that both biology and intuition need purely ahistorical functions, uncluttered by history. His admission implies that two of his key arguments for the view (cited above), are unsound. First, by his own admission, it's not the case that biologists don't refer to history; implicitly, when they talk about what's statistically-typical, they *are* talking about history. Second, regardless of whether or not intuition supports ahistorical functions, Boorse's theory doesn't. It's just not true, on Boorse's account, that if lions popped into being from an unparalleled saltation, their parts and processes would have functions. They wouldn't, since they don't have the right history (or to be more precise, they have no history at all). True, Boorse's history isn't the same *kind* of history that features in the selected effects theory, since it doesn't refer specifically to etiology, but it's still history, and so his arguments that appeal to the ahistoricity of his theory don't work.

3. The Propensity Theory

Bigelow and Pargetter (1987) also developed an influential "ahistorical" theory of function, the propensity theory. They reject the selected effects theory (and etiological accounts more generally) because the selected effects theory gets the *modality* of functions wrong. In other words, the statement, "functions are selected effects," if true, is contingently true; it might be true on the actual world, but there are possible worlds at which it's false. To illustrate the point, they ask us to consider a world that is pretty much the same as ours except that it randomly popped into being five minutes ago. On that world, they claim, there would still be functions, just no selected effects (188): "we have the intuition that the concept of biological function…[is] not thus contingent upon the acceptance of the theory of evolution by natural selection." This consideration prompts the need for an ahistorical theory.

For Bigelow and Pargetter, functions are propensities, or probabilistic dispositions. We might quibble over what exactly dispositions are, but any good definition will cite three parts: structure, environment, and behavior. Consider the solubility of salt. There is a *structure*, namely, the polar molecular structure composed of sodium and chloride; there is an *environment*, namely, water; there is a *behavior*, namely, dissolving. When we say that salt is disposed to dissolve in water, we're saying that, if you were to take this structure, and put it in this environment, it would perform this behavior.

Functions, too, are dispositions. Consider "the function of the heart is to circulate blood." For this statement to be true, there must be a structure (the heart, embedded the right way in the circulatory system), an environment (which they call the creature's *natural habitat*), and a behavior (conferring a fitness boost on the organism). If one were to put the structure in its natural habitat, it would increase the fitness of the organism (relative, I suppose, to creatures without hearts). The crucial distinction between their view and Boorse's is that in their view, a trait's function doesn't depend on actual frequencies of performance. A trait needn't have an actual track record of boosting fitness to have a function; a mere propensity will do.

This raises the thorny question of what a creature's *natural habitat* is. For they're clear that a creature's natural habitat isn't just any environment the creature happens to find itself in. Unfortunately, they refuse to define this crucial notion; instead, they brush it off as vague, but unproblematically so: "there may be room for disagreement about what counts as a creature's 'natural habitat;' but this sort of variable parameter is a common feature of many useful scientific concepts" (192). But one could at least form the suspicion that if one analyzed this unproblematically vague notion, one would find some reference to history tucked away inside of it.

This suspicion is confirmed in the very next paragraph. There, they tell us that, if a creature's environment were to change very suddenly, then "natural habitat" will still refer to the *old* environment, and not the *new* one (ibid). There's a time lag built into the very idea of a natural habitat. So, for example, if climate change melts enough Arctic ice, then, at least for a time, the polar bear's natural habitat (and by extension, the natural habitat of the trait itself, namely, their thick, water-repellant fur) is the icy habitat of yore and not the contemporary, denuded one. They take that as given, and I agree.

But why would this be? What *makes it the case* that this is true, namely, that in cases of rapid habitat change, "natural habitat," at least for a time, refers to the old environment and not the new one? What makes it true, I suspect, is that the idea of a natural habitat is an intrinsically historical notion. It's something like the environment within which the organism recently survived and thrived. And if that's not what a natural habitat is, I would like to know what it is *such that*, if a creature's actual habitat shifts suddenly, the natural habitat is still the old one. Just because a concept is vague around the edges, that doesn't exempt one from the obligation to give some sort of analysis.

Hence, I conclude that, contrary to rumor, the propensity theory is not an ahistorical theory, or not demonstrably so. But if that's right, they lose one of the main virtues of the view, which is to get the modality of functions right. To be fair, there's still a sense in which their view *is* ahistorical. What they can do, that the selected effects theorist can't, is to attribute functions to novel traits – so long as that novel trait belongs to the members of a species that has been around long enough to have a natural habitat. Suppose a gene mutation confers a benefit on an organism, say, pesticide resistance on a flour beetle. I suppose they can say that, at the very moment at which it first confers that benefit, the gene mutation has a function, namely, to make the beetle withstand a certain pesticide. This result, they claim, is "intuitively comfortable" (195). But they can say that only

because flour beetles themselves have a history, and so we can talk meaningfully about their natural habitats. Moreover, I think they'll still have a very hard time dealing with dysfunction (Neander 1991, 183), as I hope to show in the next section. Finally, I think there are good theory-neutral reasons for saying that beneficial traits, on their very first appearance, don't have functions, but rather, whatever benefit they bring is an accident. But I won't argue for that here (see OMITTED).

4. The Causal Role Theory

What about the causal role theory of function? This appears to be a purely ahistorical view. The causal role theory says, roughly, that the function of a *component* of a system consists in its contribution, in tandem with the other components, to a system-level capacity of interest (Cummins 1975; Craver 2001; Hardcastle 2002). Craver (2001; 2013) helpfully elaborates this view by specifying that the part in question must be a component of a *mechanism*. All of the basic ingredients of this theory are ahistorical: capacities, components, organization, hierarchy, interests. Even if the world were created five minutes ago, in pretty much its present form, things would still have causal role functions.

The problem enters when we think about dysfunction. Cummins (1975, 758) insisted that functions are dispositions, or capacities: "…to attribute a function to something is, in part, to attribute a disposition to it." The function of a trait *token*, then, consists in its capacity to contribute to a system-level effect. But what if the token in question, through defect or disease, loses the capacity, and so can't contribute to the system-level effect? Then, by Cummins' analysis, it doesn't have the relevant function – so it can't dysfunction either.

Causal role theorists have, by and large, been silent about how to make sense of dysfunctions from this perspective. Almost everything they've had to say on that score, however, is consistent with the following theme: a trait *token* dysfunctions when it can't do what other trait tokens generally, or typically, do to contribute to the system-level effect of interest. Consider Godfrey-Smith (1993, 200): "Although it is not always appreciated, the distinction between function and *mal*function can be made within Cummins' framework…If a token of a component of a system is not able to do whatever it is that other tokens do, that plays a distinguished role in the explanation of the capacities of the broader system, then that token component is malfunctional." Craver (2001, 72), offers the same general line: "…the ascription of a function to a malformed or broken part is derivative upon a description of how that *type* of part (X) fits into a *type* of higher-level mechanism (S). The malformed and broken part can be identified as an X by the typical properties and activities of Xs...." This is, at root, to rely on a statistical norm for making sense of dysfunction.

This account of dysfunction, like Boorse's, stumbles when it encounters the problem of pandemic dysfunction (Neander 1991). For the modification suggested above implies that, if everyone's heart seized up at once, nobody's heart would have a function anymore, so nobody's heart would be dysfunctional. The best way to solve this problem,

and perhaps the only way, is the way Boorse took, namely, to say that the function of a trait is its typical contribution to some system effect, when what's typical is assessed over a chunk of time that stretches back into the past, for at least "a lifetime or two," and perhaps "millennia." But if causal role theorists take that line, they'd have a historical theory.

Craver (2001) and Hardcastle (2002) suggest, all too fleetingly, a different way of thinking about dysfunction, one that depends not on statistics, but on our values and goals, that is, the values and goals of people who make function attributions. Craver (2001, 72) suggests that traits dysfunction when they cannot do what people *want* them to do: "the mechanistic role of the broken part only appears against the fixed backdrop of shared assumptions about a type of mechanism within which parts of this type generally (or preferably) make important contributions." The parenthetical remark alludes to a substantially new doctrine, one that demands our full concentration. It suggests that dysfunction is a mirror of human preferences and goals, of our wishing and wanting. If my heart seizes up, it's dysfunctional, since it's not doing *what I want it to do.*

Hardcastle (2002) makes remarks along similar lines. She first says that the function of a trait - what it's "supposed to do," as she puts it - depends on the goals of the scientific discipline that makes the investigation: "The teleological goal for some trait…depends upon the discipline generating the inquiry" (153). The palmomental reflex causes a chin twitch when you stroke an infant's palm; it's just an accident of cortical wiring with no deep evolutionary rationale. Still, she says, it has the *function* of indicating the state of brain development in infants, because that's how biomedical researches use it. She then says that something malfunctions just when it cannot do what it's supposed to do (152). The palmomental reflex malfunctions when it can't indicate the state of brain development. Simply put, dysfunction happens when a trait can't do what we want.

But dysfunctions cannot be reduced to preferences in any straightforward way; this is a point that's been taken for decades (e.g., Boorse 1977, 544; Wakefield 1992, 372), for reasons that scarcely need to be rehearsed. I'd prefer not to need sleep and water; I'd prefer if nobody had to go through the pain of childbirth or teething, either. But none of those things are diseases or dysfunctions. For that matter, I'd prefer if my hands were equipped with retractable adamantium claws. The fact that my hands can't do what I want them to do doesn't make them dysfunctional. If one really wanted to run with this value-centered line about dysfunction, one would *at least* have to add that, in order for a trait to dysfunction, it's not enough that it doesn't do what I prefer, but I must also have a *reasonable expectation* that it *should* act in the way that I prefer. But what could possibly ground a *reasonable expectation* that my hand (say) work in a certain way? Only this: that hands usually *do* work in the preferred way. But then we're back to statistical norms, and long historical slices of time. This value analysis of dysfunction isn't a contender to a statistical analysis; instead, the former presupposes the latter.

I've walked through three allegedly ahistorical theories of function, and shown that none of them are purely ahistorical; they're tainted with history. The conclusion will say what we should do next.

5. Conclusion

There are no ahistorical theories of function, at least among those that are usually put forward as ahistorical. The first, Boorse's goal-contribution theory, explicitly refers to what is statistically typical for a trait, where what's typical is assessed over a long historical period of time. The second, the propensity theory, refers to the creature's natural habitat, which is implicitly historical. And the third, the causal role theory, can't hope to make sense of dysfunction (or so I argue) without appealing to a statistical norm, and thereby (following Boorse) to history. *No* theory of function will give functions to the parts of swamp creatures, instant lions, or anything on worlds that are similar to ours except for being randomly generated five minutes ago. The propensity theory, at least, can give functions to novel traits as soon as those traits begin benefiting their bearers, as long as the population in which the traits emerge has been around for long enough to have something like a natural habitat. But even that theory will probably encounter problems when it comes to making sense of dysfunction, though I haven't pushed that line in any detail here.

Three immediate consequences follow from this fact. The first is that we should stop dividing up theories of function in terms of historical and ahistorical. The second is that many of the main arguments for the allegedly ahistorical theories are unsound. Third, one popular form of pluralism, which says that there are two main theories of function, corresponding to historical and ahistorical uses of "function" in biology, is untenable.

But if we can't rely on the historical/ahistorical distinction as a way of dividing up functions, how should we talk about them? I think it's best to divide them up into etiological and non-etiological (as theorists are sometimes wont to do anyway). But there's a crucial clarification in order: to say a theory is etiological isn't *just* to say it's historical. It's to say that the theory deals specifically with causal history. The theory purports to capture the sense in which, when we attribute a function to a trait, we're trying to give a causal explanation for why the trait exists. Most other theories of function are non-etiological, in that they do not purport to explain, in a causal sense of "explain," why the trait exists. But they're still historical.

There's a twist to this story. I think there *are* ahistorical theories of function. Consider that climate change is a function of deforestation, poor academic performance is a function of malnutrition, and wildlife habitat is a function of soil. These notions are *ahistorical* through and through. "Function," in this context, means little more than "effect," and perhaps (as in the last of the three examples) "helpful effect." But this tepid sense of function isn't going to sustain a distinction between function and accident, nor will it give us any sense of dysfunction. This is the sort of "function" that Bock and von Wahlert (1965, 274) were getting at when they equated functions with "all physical and chemical properties arising from [the trait's] form." It's also the sort of "function" that Neander (2017) describes in her recent discussion of "minimal functions." But the proponents of the allegedly ahistorical theories want functions to do much more than that. They are trying to capture the ordinary biological sense (or *an* ordinary biological sense)

of "function," where functions differ from accidents and sometimes things dysfunction. Unfortunately, they can't have what they want.

References

Amundson, R., and G. V. Lauder. 1994. Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy* 9: 443-469.

Bigelow, J., and Pargetter, R. 1987. Functions. *Journal of Philosophy* 84: 181-196.

Bock, W. J., and von Wahlert, G. 1965. Adaptation and the form-function complex. *Evolution* 19: 269-299.

Boorse, C. 1976. Wright on functions. *Philosophical Review* 85: 70-86.

Boorse, C. 1977. Health as a theoretical concept. *Philosophy of Science* 44: 542- 573.

Boorse, C. 2002. A rebuttal on functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M. Perlman, 63-112. Oxford: Oxford University Press.

Buller, D. J. 1998. Etiological theories of function: A geographical survey. *Biology and Philosophy* 13: 505-527.

Craver, C. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Craver, C. 2013. Functions and mechanisms: A perspectivalist view. In *Function: Selection and Mechanisms*, ed. P. Huneman, 133-158. Dordrecht: Springer.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72: 741–765.

Godfrey-Smith, P. 1993. Functions: Consensus without unity. *Pacific Philosophical Quarterly* 74: 196-208.

Hardcastle, V.G. 2002. On the normativity of functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M Perlman, 144-156. Oxford: Oxford University Press.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Neander, K. 1983. *Abnormal Psychobiology*. Dissertation, La Trobe.

Neander, K. 1991. Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.

Neander, K. 2017. Functional analysis and the species design. *Synthese* 194: 1147-1168.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47: 373–388.

Walsh, D.M. 1996. Fitness and function. *British Journal for the Philosophy of Science* 47: 553-574.

Walsh, D. M., and A. Ariew. 1996. A taxonomy of functions. *Canadian Journal of Philosophy* 26: 493-514.

Wright, L. 1973. Functions. *Philosophical Review* 82: 139-168.

What do molecular biologists mean when they say 'structure determines function'?

Gregor P. Greslehner*

University of Salzburg & ERC IDEM, ImmunoConcept, CNRS/University of Bordeaux

October 2018

Abstract

'Structure' and 'function' are both ambiguous terms. Discriminating different meanings of these terms sheds light on research and explanatory practice in molecular biology, as well as clarifying central theoretical concepts in the life sciences like the sequence–structure–function relationship and its corresponding scientific "dogmas".

The overall project is to answer three questions, primarily with respect to proteins: (1) What is structure? (2) What is function? (3) What is the relation between structure and function?

The results of addressing these questions lead to an answer to the title question, what the statement 'structure determines function' means.

*Email: gregor.greslehner@gmail.com

Keywords: philosophy of biology, molecular biology, protein structure, biological function, scientific practice

1 Introduction

'Structure' and 'function' are abundantly used terms in biological findings. Frequently, the conjunct phrase 'structure *and* function' or the directional phrase '*from* structure *to* function' is to be found, indicating that there is a special relation connecting these two concepts. The strongest form of this relation is found in the frequent statement that 'structure *determines* function'. One could easily list several hundreds of references containing such phrases. However, in order not to blow up the references section, I will refrain from doing so. Suffice it to say that biologists make highly prominent use of these concepts in describing their research—molecular biologists, in particular. In this paper, I attempt to clarify these concepts, address their relation, and discuss the role they play in molecular biology's explanatory practice. While these issues can be addressed for many different biological entities on different levels of organization, I restrict the discussion primarily to proteins.

What do biologists refer to when they use this phrase? Is there a particular scientific program or strategy behind the slogan 'structure determines function'? Despite the frequent use of this phrase and the concepts to which it refers, a rigorous analysis is missing. Thus, a philosophical clarification would be a valuable contribution to the conceptual foundations of biology. One such fundamental concept is the sequence–structure–function relationship. "The relationships between sequence, structure, biochemical function and biological role are extremely ill-defined and scant

high quality data are available to allow us to analyse them." (Sadowski and Jones, 2009, 360)

In this paper, I attempt to close this gap by developing an explication of both concepts of *structure* and *function* as they are used in biological practice and discussing which relation holds between them. The third component in this "trinity of molecular biology"—sequence—is the least in need of explication. The standard textbook view holds that sequence determines structure, and structure determines function. I will focus on the second relation.

Without reviewing the rich history of these concepts throughout biology at this point, it is worth noting that functionality and form or structure were thought to be intimately linked from early on. In the early days of biology at the macroscale, the structures had to be observed with the naked eye. Thus, the first examples about the form of bodies or their parts and their functions can be found in physiology and anatomy, for example Harvey's notion of the heart's function to pump blood. From the scale of physiology to the molecular scale, structure and function are closely related. What exactly links these two concepts? Is it a determination relation? And if so, which one is determining the other?

With the invention of microscopes and later the emergence of molecular biology, the structures and functions under consideration shifted from macroscopic entities to individual molecules. In fact, molecular biology put the three-dimensional shape of molecules center stage for explaining biological phenomena. This is the focus of this paper. In particular, the discussion will be confined to the structure and function of *proteins*—with special emphasis on the question whether the former determines the latter.

2 The ambiguity of 'structure'

In a first approximation, 'structure' and 'function' could be interpreted as the most general or neutral way of describing what molecular biologists are doing in their research and what their findings are about. These include mainly the three-dimensional shapes of molecules (or larger cellular structures) and the activities (functions) theses molecules perform in living cells, biochemical pathways, chemical reactions, or just individual steps in such mechanisms. The ultimate aim is to explain biological phenomena with molecular mechanisms, whose entities can be described in physical and chemical terms. The structure of molecules can be described in terms of physics and chemistry—function, however, is a concept that does not appear in physics or chemistry. Let's start by taking a closer look at the notion of structure.

'Structure' is an ambiguous term. Applied to proteins, there is the usual nomenclature of *primary structure* (i.e., a protein's amino acid sequence), *secondary structure* (i.e., common structural motifs like α-helices and β-sheets), *tertiary structure* (i.e., the three-dimensional shape of a single folded amino acid chain), and *quaternary structure* (i.e., the final assembly of a protein if it consists of more than one amino acid chain). Other structurally important components are post-translational modifications and prosthetic groups which are not part of its amino acid composition. All these notions of structure have in common that they are about the molecular composition and shape of a molecule. One meaning of 'structure' denotes the sequence of a polymer, the other meaning is about the three-dimensional shape of a molecule. As will be discussed below, another important ambiguity of 'structure' allows to denote the organization of an interaction network. That leaves us with three different meanings of 'structure':

(1) the sequence of a polymer, (2) the three-dimensional shape of a molecule, and (3) the network organization of several biological entities.

While meanings (2) and (3) are candidates for being functional entities, structure as sequence (1) rather relates the sequences of different polymers (DNA, RNA, and proteins) and also plays a central role in determining the three-dimensional shape of a molecule, structure (2). The primary structure of a protein is just the sequence of amino acids that are put together to form a polypeptide. This amino acid sequence is determined by the corresponding protein-coding gene, which is first transcribed into mRNA and then translated into protein by the ribosome. This scheme is known as the "central dogma of molecular biology":

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$$

The arrows might be interpreted as determination relations. The textbook view of protein structure and function proceeds as follows:

$$\text{nucleotide sequence} \rightarrow \text{amino acid sequence} \rightarrow \text{protein structure} \rightarrow \text{protein function}$$

Strong evidence supporting the claim that the three-dimensional shape of a protein is determined by the sequence of amino acids alone was provided by the experiments of Christian Anfinsen, showing that ribonuclease could, after treatment with denaturing conditions, regain its form and function (Anfinsen et al., 1961). Later, Merrifield showed that an *in vitro* synthesized sequence of amino acids can carry out the enzymatic activity of ribonuclease, thus gaining its functional form without the aid of any other cellular component (Gutte and Merrifield, 1971). From this and similar experiments, Anfinsen

built general rules of protein folding as a global energy minimum which depends solely on the sequence of amino acids (Anfinsen, 1973). This view is known as "Anfinsen's dogma".

In 1958, John Kendrew's lab determined the first actual three-dimensional form of a protein, myoglobin (Kendrew et al., 1958). The predominant technique to determine protein structures is still X-ray crystallography (Mitchell and Gronenborn, 2017). Other techniques include nuclear magnetic resonance, cryogenic electron microscopy, and atomic force microscopy. X-ray structures in particular have been supporting the view that there is a unique rigid shape—the protein's native, functional state—which would be necessary and sufficient for a protein to carry out its biological function.

To make a long story short, the relation between nucleotide sequence and amino acid sequence has been generally confirmed (although there are much more complicated mechanisms to it, e.g., splicing). However, the part concerning the protein shape and function proves to be much more problematic. That poses a challenge to what Michel Morange calls "the protein side of the central dogma" (Morange, 2006).

To get from amino acid sequence to three-dimensional structure is known as the *protein folding problem*. As the term 'problem' suggests, it poses a serious challenge and remains unsolved to this day. Even though knowledge-based techniques to predict protein structures from their sequence have become impressively sophisticated, successful, and reliable, there are good reasons to suspect that the protein problem might remain unsolved in principle—if the aim is to predict protein folding based on chemical and physical principles only.

Every two years the best prediction tools are tested in a contest, the Critical Assessment of protein Structure Prediction (CASP). Based on experimentally determined structures which are only published after the participants of the contest have

submitted their predictions, the predictions are then compared to the experimental structure. A similar contest for predicting the functions of proteins exists (Critical Assessment of Functional Annotation, CAFA), although it is much less developed. But what is function in the first place?

3 The ambiguity of 'function'

'Function' is also an ambiguous term (Millikan, 1989)—even more so than 'structure'. There is a rich history of debates surrounding different notions of function. The term 'function' has a long tradition in biology and its philosophy (Allen, 2009). Starting with Aristotle, activities in biology were interpreted to *have a purpose*, to be goal-directed (teleological). The standard example is that the heart's function is to pump blood. That the heart also produces noise is not considered to be functional. Classic accounts of function have been predominantly trying to capture the teleological aspect, for example (Wright, 1973). However, intentionality is a problematic notion in biology. In another important account, Robert Cummins (1975) stressed the importance of a component's contribution to the system in which it is contained, rather than why natural selection has favored a certain trait. Although it makes sense in evolutionary biology to have an account of function that captures the evolutionary developments, molecular biology and protein science operate with a different notion of function, i.e., mainly biochemical activity. There seem to be two entirely different questions: What is a structure doing? And how did this structure evolve to do what it does?

Arno Wouters distinguishes four notions of biological function (Wouters, 2003): (1) (mere) activity, (2) biological role, (3) biological advantage, and (4) selected effect.

The last two are issues of evolutionary biology, whereas the former two fall within the molecular biologist's domain. If function is to be determined by a molecule's three-dimensional shape or organization network, only (1) and (2) seem to be the proper reading of 'function' in this context.

Which entities have functions within living organisms? Depending on the level of organization at which one is operating, one could give a different answer: molecules, organelles, cells, tissues, organisms, individuals, populations, ecosystems. The most prevalent candidates in molecular biology are certainly DNA and proteins, although lipids and other biomolecules play important roles in life processes, too.

Traditionally, functions have been attributed to entire genes ("one gene—one enzyme hypothesis"). These views are related to the genetic determinsim view of having a gene for every trait, in which every gene has a function. However, the primary functional units inside a cell are arguably its proteins. Their biochemical activities and biological roles depend crucially on their three-dimensional shapes and network organization, respectively.

One has also to take into account more abstract functional entities, i.e., network modules. These are also called 'structures' but do not refer to the shape of molecules. Its functions ought to be considered as Wouter's second notion (biological role), rather than biochemical activity. "Current 'systems' thinking attributes primary functional significance to the collective properties of molecular networks rather than to the individual properties of component molecules" (Shapiro, 2011, 129). "[A] discrete biological function can only rarely be attributed to an individual molecule [...]. In contrast, most biological functions arise from interactions among many components." (Hartwell et al., 1999, C47). Thus, we can attribute functions as biochemical activities to

individual molecules, wheres systems functions (biological roles) are attributed to organizational structures:

> "Finding a sequence motif (e.g., a kinase domain) in a new protein sheds light on its biochemical function; similarly, finding a network motif in a new network may help explain what systems-level function the network performs, and how it performs it." (Alon, 2003, 1867)

4 Does structure determine function?

Having distinguished between three notions of 'structure' and two notions of 'function', what about the statement 'structure determines function'? Is—in any of its different readings—a certain structure necessary or sufficient for a certain function?

The common textbook view according to Anfinsen has a clear answer: "the central dogma of structural biology is that a folded protein structure is necessary for biological function" (Wright and Dyson, 1999, 322). On first glance, it might appear plausible to assume that a particular structure (understood as molecular shape) is a necessary condition for the proper function of a biological structure (i.e., its biochemical activity). Loss of function is often associated with a loss of the three-dimensional shape of individual proteins. On the other hand, to go for the "sufficient" direction, changes in structures often lead to a decrease in functionality, up to a complete loss. Many diseases for which there are known molecular causes give support to this view. Often it is alterations in the sequence of DNA that result in changed protein shapes that lead to a functionality defect of the organism, which is the definition of a "molecular disease". Alterations of a protein's three-dimensional shape, however, do not necessarily lead to

loss of function. In many cases, changes are "silent", i.e., they don't cause any alteration in phenotype. In rare events, changes might even turn out to be "improvements", which is the driving force of evolutionary development.

However, evidence has been found in the recent years that a significant portion of proteins are intrinsically unstructured in order to be functional, see for example (Forman-Kay and Mittag, 2013). Does the discovery of intrinsically unstructured proteins challenge the relation between structure and function? "[D]isorder aficionados are calling for a complete reassessment of the structure-function paradigm" (Chouard, 2011, 151). Some protein domains fold only upon binding to a suitable target. Others, however, seem to never have an ordered state at all—they remain unstructured even in their functional state.

That a high similarity in sequence does not guarantee a similarity in structure or function has been shown by the Paracelsus Challenge: "a one-time prize of $1000, to be awarded to the first individual or group that successfully transforms one globular protein's conformation into another by changing no more than half the sequence" (Rose and Creamer, 1994, 3). One recent answer to this challenge resulted in the synthesis of two proteins which have 88% sequence identity but a different structure and a different function (Alexander et al., 2007).

Contrary to the view described above, the generalization that a stable three-dimensional structure is necessary or sufficient for a particular function does not hold. It remains true, however, that there is an intimate correlation between structure and function. Prediction tools based on this view are a powerful tool. An attempt to systematically predict the structure and function of proteins based on their amino acid sequence can be found, for example, in (Roy et al., 2010).

To complicate the picture, codon usage is also important: Zhou et al. (2013) have shown that the FRQ protein, which is involved in the circadian clock, is using non-optimal codons, thus translation speed is not optimal. After experimentally optimizing codon usage, the resulting protein—which has the exact same amino acid sequence—folds differently and is no longer functional. This shows that amino acid sequence by itself is not sufficient to determine the three-dimensional structure, let alone its function. In addition to the correct sequence, the folding process has to take place in a certain way which is influenced by the usage of codons and thus the availability of tRNAs, which influences the speed at which the ribosome can proceed translation. Usage of non-optimal codons gives the nascent polypeptide chain some time for the segments that have already been translated to fold in a certain conformation. If translation is too fast, certain intermediate folds which are necessary to reach the final functional conformation can be lost.

Another idea to keep in mind is that evolution operates pragmatically: structures are not the target of selection, functions are. Structures are being re-used for novel functions—there are many biological examples.

If structure does not *determine* function, if a particular structure (in any of its three meanings) is neither necessary nor sufficient for a particular function (in any of its two meanings), may there be another way in which structure and function are related? Perhaps there is a less stringent relationship? I will argue for a supervenience relation (McLaughlin and Bennett, 2018). But before developing this account, we need to clarify which notions of 'structure' and 'function' to use to capture actual scientific practice in molecular biology.

In order to speak about biological functions, a reglemented vocabulary is needed.

The most successful of these is gene ontology (GO) (Ashburner et al., 2000). Fascinating correlation analysis between three-dimensional protein structures from the Protein Data Bank (PDB) and GO terms can be found, for example, in (Hvidsten et al., 2009) and (Pal and Eisenberg, 2005).

According to the textbook picture, there is a linear chain of determination, leading from nucleotide sequences in the DNA via transcription to the nucleotide sequence of RNA, which leads via translation to the amino acid sequence of proteins. The sequence of amino acids, in turn, determines the three-dimensional structure of the protein, whose function, again, is determined by its structure. Given transitivity of this determination relation, one would only need to know the genomic sequence in order to have a complete picture ("blue print") of the functional organism. That is the "holy grail of molecular biology". And like the quest for the holy grail, it is doomed to fail. A strict determination relation does not even hold between the individual pairs.

The reason why the simplified scheme above is still part of the current research "paradigm" lies, on the one hand, in its scientific success: genomics and proteomics have provided unimaginable insights. On the other hand, it fits the mechanistic, reductionistic narrative that has been fashionable in molecular biology. Today, systems biology claims to provide a "holistic" alternative (Green, 2017).

But even without such a strict determination relation between structure and function, both concepts are central to explaining molecular mechanisms in research practice.

In order to understand why molecular biologists explain mechanisms with reference to structure and function, we need to understand what these concepts denote. In a first approximation, molecular biologists analyze a phenomenon by identifying its components that are responsible for the phenomenon in question. These components are the

structures that perform certain biochemical activities, which collectively bring about the phenomenon (biological role). The way in which these entities and their activities are organized is a different meaning of 'structure' which is as important in a mechanistic explanation as individual molecular structures are.

> "Despite the lack of an overarching theory, a Newtonian or quantum mechanics of its very own, molecular biology has become a unifying discipline in virtue of the powers of its techniques, its ability to extrapolate from the molecular to higher levels, and its synthesis of problems of form and function at the molecular level. This synthesis of form and function is a central, ill-understood, and historically important feature of molecular biology." (Burian, 1996, 68)

The ambiguity of the terms 'structure' and 'function' might be useful, for it can be applied to a broad variety of biological research strategies and activities. But, on the other hand, using the term same for different things causes confusion, and the use of metaphorical language might be obscuring certain features and difficulties with this approach.

More recent and thriving approaches in the life sciences have moved beyond the idea that there is a determination relation between structure and function and that by knowing the structure of a protein one could predict its biological function. Today's research in molecular biology is more centered around the *organizational structure* of biological mechanisms. In this way, the ambiguity of the term 'structure' suits to uphold the research slogan, since it can also be applied in a broader sense here than just molecular shapes. The organization of biological systems is the domain of the relatively

new discipline systems biology.

The three-dimensional shape is often a detail that does not contribute to the understanding of a mechanism, but to the contrary would only confuse the mechanistic picture which requires a certain level of abstraction in order to be comprehensive.

But still, how exactly do we get from molecular structures and their (structured) activities to biochemical activities and biological functions? That there might not exist a straightforward mapping from molecular shapes to their biochemical and biological function had been anticipated in the early days of molecular biology:

> "It [molecular biology] is concerned particularly with the *forms* of biological molecules, and with the evolution, exploitation and ramification of these forms in the ascent to higher and higher levels of organization. Molecular biology is predominantly three-dimensional and structural—which does not mean, however, that it is merely a refinement of morphology. It must of necessity enquire at the same time into genesis and function." (Astbury, 1952, 3, original emphasis)

Taking up Francis Crick's remark that "folding is simply a function of the order of the amino acids" (Crick 1958, 144), Morange comments that it is "obviously not a *simple* function" (Morange, 2006, 522). And he observes a semantic change in the meaning of 'function':

> "For Francis Crick, function meant the application of simple rules and principles. For specialists today, function is the result of a complex evolution [...] This shift in the meaning of a word is more than anecdotal. It reflects an active ongoing transformation of biology [...] The mechanistic models of

> molecular biology are no longer considered sufficient to explain the structures and functions of organisms. They have to be complemented and allied with evolutionary explanations" (Morange, 2006, 522).

In order to explain biological phenomena, there is no determination relation that would allow us to track everything down to the chemical and physical properties of proteins, let alone the nucleotide sequences of DNA. Of course, all these issues are relevant to the topic of reduction:

> "if [...] regulatory networks turn out to be crucial to explaining development (and evolution [...]), the reductionist interpretation *may* be in trouble. If network-based explanations are ubiquitous, it is quite likely that what will often bear the explanatory weight in such explanations is the topology of the network rather than the specific entities of which it is composed. [...] How topological an explanation is becomes a matter of degree: the more an explanation depends on individual properties of a vertex, the closer an explanation comes to traditional reduction. The components matter more than the structure. Conversely, the more an explanation is independent of individual properties of a vertex, the less reductionist it becomes." (Sarkar, 2008, 68, original emphasis)

5 Conclusion

Both terms, 'structure' and 'function', are highly ambiguous. So is the widely used conjunct phrase of 'structure and function' that is ubiquitous in biology, as well as the

even stonger claim 'structure determines function'. Perhaps this is why it can be used in many different contexts and for many different explanatory aims in biology. Although providing a certain framework of generality, I argue that a clarification of these concepts is beneficial—for conceptual and philosophical considerations, as well as for the way biologists think about the grand schemes like the "central dogma". Ideally, such an account would also have practical implications and benefit current biological research.

To sum up the results of my analysis, in molecular biology's explanatory practice, 'structure' may refer to:

1. the sequence of polymers,
2. the three-dimensional shape of molecules (or their parts), and
3. the way biological entities are organized.

Of course, different aspects of this distinction play different roles in the explanatory practice with respect to molecular mechanisms. The detailed shape of the interacting molecules is neither necessary nor sufficient for understanding its activities (although correlations are valuable prediction tools before doing experiments in the lab).

The ambiguity of the term 'function' depends on whether the explanation aims at answering the question how a mechanism works or how it came to work that way. Even in the first case one has to distinguish between:

1. the biochemical activity of individual components, and
2. the biological role of network structures.

Whereas biochemical activities of proteins can often be successfully predicted by homology modeling from known molecular shapes, the biological role is rarely an

intrinsic property of an isolated molecule. Rather, the biological role is the mechanistic result of an interaction network of several dynamically interacting molecules.

By comparing the combinatorial possibilities of the different meanings of 'structure' and 'function', a determination relation does not hold between any of them. Instead, I propose a supervenience relation: between the three-dimensional shapes of protein domains and their biochemical activities, and between interaction networks and their biological role. According to my analysis, this is what molecular biologist mean when they say 'structure determines function'.

References

Alexander, P. A., Y. He, Y. Chen, J. Orban, and P. N. Bryan (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences 104*(29), 11963–11968. doi:10.1073/pnas.0700922104.

Allen, C. (2009). Teleological notions in biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009 ed.). `http://plato.stanford.edu/archives/win2009/entries/teleology-biology/`.

Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science 301*(5641), 1866–1867. doi:10.1126/science.1089072.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science 181*(4096), 223–230. doi:10.1126/science.181.4096.223.

Anfinsen, C. B., E. Haber, M. Sela, and F. H. White, Jr (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences 47*(9), 1309–1314.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. *Nature Genetics 25*, 25–29. doi:10.1038/75556.

Astbury, W. T. (1952). Adventures in molecular biology. In *The Harvey Lectures. Delivered under the auspices of the Harvey Society of New York. 1950–51*, pp. 3–44. Charles C Thomas.

Burian, R. M. (1996). Underappreciated pathways toward molecular genetics as illustrated by Jean Brachet's cytochemical embryology. In S. Sarkar (Ed.), *The Philosophy and History of Molecular Biology: New Perspectives*, pp. 67–85. Kluwer Academic Publishers.

Chouard, T. (2011). Breaking the protein rules. *Nature 471*, 151–153. doi:10.1038/471151a.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy 72*(20), 741–765. doi:10.2307/2024640.

Forman-Kay, J. D. and T. Mittag (2013). From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure 21*(9), 1492–1499. doi:10.1016/j.str.2013.08.001.

Green, S. (2017). Philosophy of systems and synthetic biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 ed.). `https://plato.stanford.edu/archives/sum2017/entries/systems-synthetic-biology/`.

Gutte, B. and R. B. Merrifield (1971). The synthesis of ribonuclease A. *Journal of Biological Chemistry 246*, 1922–1941.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray (1999). From molecular to modular cell biology. *Nature 402*(6761 Suppl.), C47–C52. doi:10.1038/35011540.

Hvidsten, T. R., A. Lægreid, A. Kryshtafovych, G. Andersson, K. Fidelis, and J. Komorowski (2009). A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLoS ONE 4*(7), e6266. doi:10.1371/journal.pone.0006266.

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, and H. Wyckoff (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature 181*(4610), 662–666. doi:10.1038/181662a0.

McLaughlin, B. and K. Bennett (2018). Supervenience. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). `https://plato.stanford.edu/archives/spr2018/entries/supervenience/`.

Millikan, R. G. (1989). An ambiguity in the notion "function". *Biology and Philosophy 4*, 172–176. doi:10.1007/BF00127747.

Mitchell, S. D. and A. M. Gronenborn (2017). After fifty years, why are protein X-ray

crystallographers still in business? *The British Journal for the Philosophy of Science 68*(31), 703–723. doi:10.1093/bjps/axv051.

Morange, M. (2006). The protein side of the central dogma: Permanence and change. *History and Philosophy of the Life Sciences 28*(4), 513–524.

Pal, D. and D. Eisenberg (2005). Inference of protein function from protein structure. *Structure 13*, 121–130. doi:10.1016/j.str.2004.10.015.

Rose, G. D. and T. P. Creamer (1994). Protein folding: Predicting predicting. *PROTEINS: Structure, Function, and Genetics 19*, 1–3.

Roy, A., A. Kucukural, and Y. Zhang (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols 5*(4), 725–738. doi:10.1038/nprot.2010.5.

Sadowski, M. and D. T. Jones (2009). The sequence–structure relationship and protein function prediction. *Current Opinion in Structural Biology 19*, 357–362. doi:10.1016/j.sbi.2009.03.008.

Sarkar, S. (2008). Genomics, proteomics, and beyond. In S. Sarkar and A. Plutynski (Eds.), *A Companion to the Philosophy of Biology*, pp. 58–73. Blackwell Publishing Ltd.

Shapiro, J. A. (2011). *Evolution: a view from the 21st century.* FT Press Science.

Wouters, A. G. (2003). Four notions of biological function. *Studies in History and Philosophy of Biological and Biomedical Sciences 34*(4), 633–668. doi:10.1016/j.shpsc.2003.09.006.

Wright, L. (1973). Functions. *The Philosophical Review 82*(2), 139–168. doi:10.2307/2183766.

Wright, P. E. and H. J. Dyson (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology 293*, 321–331. doi:10.1006/jmbi.1999.3110.

Zhou, M., J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. Sachs, and Y. Liu (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature 495*, 111–115. doi:10.1038/nature11833.

Seattle, WA; 1-4 November 2018 -299-

Is Peer Review a Good Idea?*

Remco Heesen[†‡] Liam Kofi Bright[§]

September 19, 2018

Abstract

Pre-publication peer review should be abolished. We consider the effects that such a change will have on the social structure of science, paying particular attention to the changed incentive structure and the likely effects on the behavior of individual scientists. We evaluate these changes from the perspective of epistemic consequentialism. We find that where the effects of abolishing pre-publication peer review can be evaluated with a reasonable level of confidence based on presently available evidence, they are either positive or neutral. We conclude that on present evidence abolishing peer review weakly dominates the status quo.

*Both authors contributed equally. Thanks to Justin Bruner, Adrian Currie, Cailin O'Connor, and Jan-Willem Romeijn for valuable comments. RH was supported by an Early Career Fellowship from the Leverhulme Trust and the Isaac Newton Trust. LKB was supported by NSF grant SES 1254291.

[†]Department of Philosophy, School of Humanities, University of Western Australia, Crawley, WA 6009, Australia. Email: remco.heesen@uwa.edu.au.

[‡]Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK.

[§]Department of Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: liamkbright@gmail.com.

1 Introduction

Peer review plays a central role in contemporary academic life. It sits at the critical juncture where scientific work is accepted for publication or rejected. This is particularly clear when the results of scientific work are communicated to non-scientists, e.g., by journalists. The question "Has this been peer reviewed?" is commonly asked, and a positive answer is frequently taken to be a necessary and sufficient condition for the results to be considered serious science.

Given these circumstances, one might expect peer review to be an important topic in the philosophy of science as well. Peer review should arguably play a more prominent role in the debate about demarcation criteria (what separates science from other human pursuits?), as it seems to be used in practice exactly to differentiate scientific knowledge from other claims to knowledge, at least by journalists. Yet as far as we know, social-procedural accounts of science, like the one found in Longino (1990), remain in the minority and usually do not place great emphasis on peer review in particular. Aside from this particular debate, there are normative questions about the proper epistemic role of peer review and more practical questions about the extent to which it manages to fulfill them, all of which should interest philosophers of science.

But there has been surprisingly little work on peer review by philosophers of science. Most of what exists has focused on the role of biases in peer review, see for example Saul (2013, §2.1), Lee et al. (2013), Jukola (2017), Katzav and Vaesen (2017), and Heesen (2018). We are not aware of any philosophical discussion of the strengths and weaknesses of peer review as such (the above examples presuppose its overall legitimacy by discussing its implementation). Some work along these lines does exist outside of philosophy, either in the form of opinion pieces (Gowers 2017) or occasionally full-length articles (Smith 2006). Such work tends to be vague about the normative standard against which peer review or its alternatives are to be

evaluated, something we aim to remedy in section 2.

Here we bring together the work of philosophers of science (especially social epistemologists of science) who have written about the strengths and weaknesses of various aspects of the social structure of science and empirical work about the effects of peer review. We argue that where philosophers of science have claimed the social structure of science works well, their arguments tend to rely on things other than peer review, and that where specific benefits have been claimed for peer review, empirical research has so far failed to bear these out. Comparing this to the downsides of peer review, most prominently the massive amount of time and resources tied up in it, we conclude that we might be better off abolishing peer review.

Some brief clarifications. Our target is pre-publication peer review, that is the review of a manuscript intended for publication, where publication is withheld until one or more editors deem the manuscript to have successfully passed peer review. We set aside other uses of peer review (e.g., of grant proposals or conference abstracts) and we explicitly leave room for post-publication peer review, where manuscripts are published before review. Because of this last point, some readers may think that our terminology ('abolishing pre-publication peer review') suggests a more dramatic change than what we actually advocate. We invite such readers to substitute in their preferred terminology. We should also clarify that we use 'science' in a broad sense to include the natural sciences, the social sciences, and the humanities.

The overall structure of our argument is as follows. We think there are a number of clear benefits to abolishing pre-publication peer review. In contrast, while various benefits of the existing system (downsides of abolishing peer review) have been suggested, we do not think there exist any that have clear empirical support. Insofar as empirical research exists, it is ambiguous in some cases, and speaks relatively clearly against the claimed benefit of the existing system in others. While we admit to a number of cases where the evidence is ambiguous or simply lacking (see especially section 5), we claim

that the present state of the evidence suggests that abolishing pre-publication peer review would lead to a Pareto improvement: each factor considered is either neutral or favors our proposal.

Our primary aim here is to evaluate the current system, but we believe that is only really possible by comparing it to an alternative. We are not claiming that the proposal we put forward is the best of all possible alternatives. It has been constructed to be a system which could constitute a Pareto improvement over the current system. Given that it has not actually been implemented yet, we cannot guarantee it would work as advertised or what empirical properties it would have. But in offering a relatively specific alternative, we hope to get people thinking about real change, which pointing out problems with the present system has so far failed to do.

Even despite this proviso, we realize that ours is a strong claim, and our proposal a large change to the social structure of science. It is therefore important to highlight that our central claim concerns the balance of presently available evidence. We are not further claiming that the matter is so conclusively settled as to render further research superfluous or wasteful. On the contrary, we think there are a number of points in our argument where the presently available evidence is severely limited, and we take the calls for further empirical research we make in those places to be just as important a part of the upshot of our paper as our positive proposal. We hope, therefore, that even a skeptical reader will read on; if not to be convinced of the need of abolishing pre-publication peer review, then at least to see where in our view their future research efforts should concentrate if they are to shore up pre-publication peer review's claims to good epistemic standing.

2 Setting the Stage

The purpose of peer review is usually construed in terms of quality control. For example, Katzav and Vaesen (2017, 6) write "The epistemic role of peer

review is assessing the quality of research", and this seems to be a common sentiment per, e.g., Eisenhart (2002, 241) and Jukola (2017, 125). But how well does peer review succeed in its purpose of quality control? The empirical evidence (reviewed below) is mixed at best. As one prominent critic puts it, "we have little evidence on the effectiveness of peer review, but we have considerable evidence on its defects" (Smith 2006, 179).

Peer review's limited effectiveness would perhaps not be a problem if it required little time and effort from scientists. But in fact the opposite is true. Going from a manuscript to a published paper involves many hours of reviewing work by the assigned peer reviewers and a significant time investment from the editor handling the submission. The editor and reviewers are all scientists themselves, so the epistemic opportunity cost of their reviewing work is significant: instead of reviewing, they could be doing more science.

Given these two facts—high (epistemic) costs and unclear benefits—we raise the question whether it might be better to abolish pre-publication peer review. In the following we provide our own survey and assessment of the evidence that bears on this question. Our conclusions are not sympathetic to peer review. However, we encourage any proponents of peer review to give their own assessment. We only ask that any benefits claimed for peer review are backed up by empirical research, and that they are epistemic benefits, i.e., we ask for empirical evidence that peer review makes for better science on science's own terms.

We take the status quo to be as follows. The vast majority of scientific work is shared through journal publications, and the vast majority of journals uses some form of pre-publication peer review. Ordinarily this means that an editor assigns one to three peers (scientists whose expertise intersects the topic of the submission), who provide a report and/or verdict on the submission's suitability for publication. The peer reviews feed into the final judgment: the submission is accepted or rejected with or without revisions.

Our proposal is to abolish pre-publication peer review. Scientists them-

selves will decide when their work is ready for sharing. When this happens, they publish their work online on something that looks like a preprint archive (although the term "preprint" would not be appropriate under our proposal). Authors can subsequently publish updated versions that reply to questions and comments from other scientists, which may have been provided publicly or privately. Most journals will probably cease to exist, but the business of those that continue will be to create curated collections of previously published articles. Their process for creating these collections will presumably still involve peer review, but now of the post-publication variety.

Our proposal is in line with how certain parts of mathematics and physics already work: uploading a paper to arXiv is considered publishing it for most purposes, with journal peer review and publication happening almost as an afterthought (Gowers 2017). Indeed, journal publication can function as something like a prize, accruing glory to the scientist who achieves it but doing little to actually help spread or diffuse the idea beyond calling attention to something that has already been made public elsewhere. We are not aware of any detailed comparative studies of the effects these changes have had in those fields, so we will not rest any significant part of our argument on this case. But for those who worry that science will immediately and irrevocably fall apart without peer review, we point out that this does not appear to have happened in the relevant parts of mathematics and physics.

In the remainder of this paper we break down the consequences of our proposal. Our strategy here is to focus on a large number (hopefully all) aspects of the social structure of science that will be affected. In particular, the reader may already have a particular objection against our proposal in mind. We encourage such a reader to skip ahead to the section where this objection is discussed before reading the rest of the paper.

For example, one reader may think that peer review as currently practiced is important because it forces scientists to read and review each other's work, and without peer review they will spend less time on such tasks. This

is discussed in section 3.2. Another reader may worry that without peer review and the journal publications that go with them it will be more difficult to evaluate scientists for hiring or promotion (section 3.5). Yet another reader may be concerned about losing peer review's ability to prevent work of little merit from being published, or at least to sort papers into journals by epistemic merit so scientists can easily find good work (section 4.1). A fourth reader might think peer review plays an important role in detecting fraud or other scientific malpractice (section 4.2). A fifth reader may think the guarantee provided to outsiders when something has been peer reviewed is an important reason to preserve the status quo (section 5.1). And a sixth reader may want to point out that anonymized peer review gives relatively unknown scientists a chance at an audience by publishing in a prestigious journal, whereas on our proposal perhaps only antecedently prominent scientists will have their work read and engaged with (section 5.2).

Other aspects of the social structure of science that will be considered: whether and when scientists share their work (section 3.1), how many papers are published by women or men (section 3.3), library resources (section 3.4), the power of editors as gatekeepers (section 3.6), science's susceptibility to fads and fashions (section 4.3), and ways to get credit for scientific work other than through journal publications (section 4.4). In each case we evaluate whether the net effects of our proposal on that aspect can be expected to be positive. To tip our hand: aspects where we will claim a benefit are gathered in section 3, aspects where we expect little or no change are in section 4, and aspects that we consider neutral due to a present lack of evidence are in section 5.

In making these evaluations, we commit to a kind of epistemic consequentialism (cf. Goldman 1999). One may think of what we are doing as roughly analogous to the utilitarian principle, where for each issue our yardstick is whether pre-publication peer review shall generate the greatest amount of knowledge produced in the least amount of time. More specifically, we con-

sider changes in the incentive structure and expected behaviour of scientists, as well as other changes that would result from abolishing pre-publication peer review. We evaluate these changes in terms of their expected effect on the ability of the scientific community to produce scientific knowledge in an efficient manner. Working out in detail what such an epistemic consequentialism would look like would be very complicated, and we do not attempt the task here. For most of the issues we consider, we think that the calculus is sufficiently clear that fine details do not matter. Where it is unclear (the issues discussed in section 5) we think this results from ignorance of empirical facts about the likely effect of policies, rather than conceptual unclarity in the evaluative metric. So we do not need to use our consequentialist yardstick to settle any difficult tradeoffs. All we need for our purposes is to make it clear that we are evaluating the peer review system by how well it does in incentivizing efficient knowledge production.

What do we mean by the incentive structure of science, mentioned in the previous paragraph? This addresses the motivations of scientists. Scientists are rewarded for their contributions with credit, i.e., with recognition from their peers as expressed through such things as awards, citations, and prestigious publications (Merton 1957, Hull 1988, Zollman 2018). Scientific careers are largely built on the reputations scientists acquire in this way (Latour and Woolgar 1986, chapter 5). As a result, scientists engage in behaviors that improve their chances of credit (Merton 1969, Dasgupta and David 1994, Zollman 2018).

While individual scientists may be motivated by credit to different degrees (curiosity, the thrill of discovery, and philanthropic goals are important motivations for many as well), the effect on careers means that credit-maximizing behavior is to some extent selected for. Thus we think it important to ensure that our proposal does not negatively affect the incentives currently in place for scientists to work effectively and efficiently.

3 Benefits of Abolishing Peer Review

3.1 Sharing Scientific Results

An important feature of (academic) science is that there is a norm of sharing one's findings with the scientific community. This has been referred to as the communist norm (Merton 1942). In recent surveys, scientists by and large confirm both the normative force of the communist norm and their actual compliance (Louis et al. 2002, Macfarlane and Cheng 2008, Anderson et al. 2010). This norm is epistemically beneficial to the scientific community, as it prevents scientists from needlessly duplicating each other's work.

Will abolishing peer review affect this practice? In order to answer this question, we need to know what motivates scientists to comply with the communist norm, that is to share their work. On the one hand there is the feeling that they ought to share generated by the existence of the norm itself. There is no reason to expect this to be changed by abolishing peer review.

On the other hand there is the motivation generated by the desire for credit. According to the priority rule, the first scientist to publish a particular discovery gets the credit for it (Merton 1957, Dasgupta and David 1994, Strevens 2003). So a scientist who wants to get credit for her discoveries has an incentive to publish them as quickly as possible, in order to maximize her chances of being first. Recent work suggests that this applies even in the case of smaller, intermediate discoveries (Boyer 2014, Strevens 2017, Heesen 2017b). All of this helps motivate scientists to share their work.

If peer review were to be abolished, the communist norm and the priority rule would still be in effect, so scientists would still be motivated to share their work as quickly as possible. However, the following change would occur.

In the absence of pre-publication peer review, scientists would be able to share their discoveries more quickly. In the current system, peer review can hold up publication for significant amounts of time, especially in the case of fields with high rejection rates or long turnaround times. During this time,

other scientists cannot build on the work and may spend their time needlessly duplicating the work. Cutting out this lag by letting scientists publish their own work when they think it is ready will speed up scientific progress. While being faster is not always better (it may increase the risk of error, cf. Heesen 2017c), in this case delays in publication are reduced without any reduction in the time spent on the scientific work itself.

To some extent this already occurs. Scientists, especially well-connected scientists, already share preprints that make the community aware of their work in advance of publication. For people who regularly do this, practically speaking little would change upon adopting the system we advocate. However, our proposal turns pre-journal-publication dispersal of work from a privilege of a well-connected few into the norm for everyone.

On this point, then, abolishing peer review is a net positive, as scientists will still be incentivized to share their work as soon as possible, but the delays associated with pre-publication peer review are removed.

3.2 Time Allocation

The current system restricts the way scientists are allowed to spend their time. For each paper submitted to a journal, a number of scientists are conscripted into reviewing it, and at least one editor has to spend time on that paper as well.

On our proposal, scientists would be free to choose how much of their time to spend reading and reviewing others' work as compared to other scientific activities. Some scientists would spend less time reviewing, some scientists would spend more, and some would spend exactly as much as under the current system.

For scientists in the latter category our proposal makes no difference, while for scientists in the other two categories our proposal represents a net improvement of how they spend their time, at least in their own judgments. We think people are the best judges of how to use their own time and labor.

We thus trust scientists' decisions in these regards, and welcome changes that would render many scientists' choices about how to allocate their own labor independent of the preferences of the relatively small number of editorial gatekeepers.

So we assume that scientists are well-placed to judge how best to use their own abilities to meet the community's epistemic needs. We claim, moreover, that the reward structure of science is set up so as to make it in their interest to do so: the credit economy incentivizes scientists to spend their time on pursuits the epistemic value of which will be recognized by the community (Zollman 2018). Hence freeing up the way scientists allocate their time leads to net epistemic benefits to the scientific community.

One might object that journals perform a useful epistemic sorting role, telling scientists what is worth spending their time on. We will address these concerns in section 4.1.

One might think that this would lead scientists to spend significantly less time reading and reviewing others' work. If this is right, we still think it would be an overall improvement for the reasons mentioned above. But we also want to point out that this is not as obvious a consequence as it may seem. Here are two reasons to expect scientists to spend as much time or more reading and reviewing on our proposal. First, for many scientists reading and reviewing are intrinsically valuable and can help their own research. Second, the current system provides no particular incentive to read and review either: scientists agree to review only because they independently want to or because they feel an obligation to the research community. While no one scientist is conscripted, at the group level editors are going to keep going until they find someone. This can amount to picking whomever is most weak-willed or under some extra-epistemic social pressure. It is not obvious that this way of deciding who does the reviewing has much to recommend it. Any rewards that exist for reviewing will still exist on our proposal, and may be amplified by the possibility of making post-publication reviews public.

3.3 Gender Skew in Publications

Male scientists publish more, on average, than female scientists, a phenomenon known as the productivity puzzle or productivity gap (Zuckerman and Cole 1975, Valian 1999, Prpić 2002, Etzkowitz et al. 2008). Several explanations have been suggested, none of which are entirely satisfactory (see especially Etzkowitz et al. 2008, 409–412). Two of these explanations that are relevant to our concerns here are the direct effects of gender bias and the indirect effects of the expectation of gender bias.

There is some evidence of gender bias in peer review, although this is not unambiguous (see Lee et al. 2013, 7–8, Lee 2016, and references therein). Insofar as there is gender bias—in the sense of women's work being judged more negatively by peer reviewers—abolishing peer review will remove this and help level the playing field for men and women. We expect positive epistemic consequences from the removal of these arbitrarily different standards.

While the evidence of gender bias in peer review is not entirely clear-cut, there is good evidence that women *expect* to face gender bias in peer review (see Lee 2016, Bright 2017b, Hengel 2018, and references therein). In an effort to overcome this perceived bias, women tend to hold their own work to higher standards. Hengel (2018) provides evidence that women spend more time correcting stylistic aspects of their paper during peer review, presumably due to higher expectations of scrutiny on such apparently superficial elements of their work. On the plausible assumption that if women have higher standards for each paper they will produce fewer papers overall, this means that the mere expectation of gender bias can contribute to the productivity gap.

After abolishing peer review both women and men will hold their work primarily to their own individual standards of quality, and secondarily to their expectations of the response of the entire scientific community, but not to their expectations of the opinion of a small arbitrary group of gatekeepers. We do not know whether this will lead the women to behave more like the men (producing more papers) or the men to behave more like the women

(holding individual papers to a higher standard of quality). However, in line with our view above that scientists are well-placed to judge how best to spend their own time, we take it that any resulting change in behavior will be a net epistemic positive.

3.4 Library Resources

Journal subscription fees currently take up a large amount of library resources (RIN 2008, Van Noorden 2013). To summarize some key figures from the 2008 report: research libraries in the UK spent between £208,000 and £1,386,000 on journal subscriptions annually (and that was a decade ago, with subscriptions having risen substantially since). The cost for publishing and distributing a paper was estimated to be about £4,000, or about £6.4 billion per year in total. Savings from moving to author-paid open access were estimated at £561 million, about half of which would directly benefit libraries.

On our proposal, this is replaced by the cost of maintaining one or more online archives of scientific publications. The example of existing large preprint archives like arXiv and bioRxiv suggests that maintaining such archives can be done at a fraction of the cost currently spent on journal subscription fees. As a rough guideline, Van Noorden (2013) estimates maintenance costs of arXiv at just $10 per article. So our proposal involves significant savings on library resources, which could be used to expand collections, retain more or better trained staff, or other purposes that would be of epistemic benefit to the scientific community.

Two additional effects should be considered in relation to this. First, the fact that the online archive will be open access means that scientific publications will be available to everyone, not just to those with a library subscription or some other form of access to for-profit scientific journals.

Second, the fact that any value added by for-profit journals would be taken away. The two tasks currently carried out by journals that could

plausibly be supposed to add value to scientific publications are peer review and copy-editing (Van Noorden 2013). It is the purpose of all other sections of this paper to argue that peer review does not in fact (provably) add value, so we set that aside. This leaves copy-editing. We propose that libraries use some of the funds freed up from journal subscriptions to employ some copy-editors. Each university library would make copy-editing services available to the scientists employed at that university. We contend that, after paying for the maintenance of an online archive and a team of copy-editors, under our proposal libraries would still end up with more resources for other pursuits than under the current system.

We note that this particular advantage of our proposal is a bit more historically contingent than the others. There seems to be no particular reason why pre-publication peer review has to be implemented through for-profit journals, and if the open access movement has its way we might be able to free up these library resources without abolishing pre-publication peer review. But our proposal also achieves this goal, and so we count it as an advantage relative to the system as it is currently actually implemented.

3.5 Scientific Careers

The 'publish or perish' culture in science has been widely noted (e.g., Fanelli 2010). Universities judge the research productivity of scientists through their publications in (peer reviewed) journals, with some focusing more on 'quantity' (counting publications) and others on 'quality' (publishing in prestigious journals). Scientific journals and the system of pre-publication peer review thus play an important role in shaping scientific careers. What will become of this if peer review is abolished?

We note first that the 'publish or perish' culture is a subset of a larger system which we discussed above: the credit economy. Publishing in a journal is one way to receive credit for one's work, but there are others, most prominently citations and awards. Scientific careers depend on all of these,

with different institutions weighting quantity of publications, quality of publications, citation metrics, and awards and other honors differently.

Any of these types of credit represents some kind of recognition of the scholarly contributions of the scientist by her peers. But here we distinguish two types of credit, which we will call short-run credit and long-run credit. Getting a paper through peer review yields a certain amount of credit: more for more prestigious journals, less for less prestigious ones. But this is short-run credit in the following sense. The editor and the peer reviewers judge the technical adequacy and the potential impact of the paper, shortly after it is written. Their judgment is essentially a prediction of how much uptake the paper is likely to receive in the scientific community.

In contrast, citations (as well as awards, prizes, inclusion in anthologies or textbooks, etc.) represent long-run credit. They *are* the uptake the paper receives in the scientific community. Long-run credit is both a more considered opinion of the scientific importance of the paper and a more democratic one (citations can be made by anyone, and awards usually reflect a consensus in the scientific community, whereas peer review is normally done by up to three individuals). So long-run credit reflects a more direct and better estimate of the real epistemic value of a contribution to science.

So what would the effect of our proposal be? For better or worse, our proposal does not make it impossible for universities to use metrics to judge research productivity. While journal rankings and impact factors would disappear, citation metrics for individual scientists and papers would still be available. This may mean that universities stop judging their scientists based on the impact factors of the journals they publish in and start judging them on the actual citation impact of their papers. More generally, our proposal will decrease or remove the role of short-run credit in shaping career outcomes and increase the role of long-run credit, which we take to be a better measure of scientific importance. So we think this is an improvement on the status quo.

What about junior hires and related career decisions, where long-run credit may be absent or minimal? If abolishing peer review means completely getting rid of journals and the associated prestige rankings, this robs hiring departments of some information regarding the scientific importance of candidates' work. If this means those on the hiring side need to read and form an opinion of candidates' work for themselves, we do not think that is a bad thing. This would of course take time, but if journals and peer review are completely abolished, that just means the time spent reviewing the paper is transferred to the people considering hiring the scientist, which again, we do not think is a bad thing. In fact, since very few academics are on a hiring committee year after year, whereas referee requests are a constant feature while one is in the community, we think that even this added burden when hiring might still be a net time-saver for academics.

But it does not have to be that way. We never said journals and peer review have to be completely abolished—our proposal in section 2 explicitly suggests journal issues may still appear, but as curated collections of articles based on post-publication peer review. So short-run credit based on journal prestige need not disappear. It need not even be slower as there is no particular reason post-publication peer review needs to take longer than pre-publication peer review. But there is the added advantage that the paper is already published while it undergoes peer review, so the wider community outside the assigned reviewers also has a chance to respond before it is included in a journal.

3.6 The Power of Gatekeepers

The discussion immediately above touched on another effect, one that we think is worth bringing out as a benefit of our proposal in its own right. As mentioned our proposal suggests that in evaluating the importance of scientific work we decrease our reliance on short-run credit (journal prestige), with a corresponding increase in long-run credit (citations, among other things).

This means that the overall credit associated with a particular paper depends less on the judgments made by an editor and a small number of reviewers, and more on its actual uptake in the larger scientific community.

Editors in particular currently play a large role in determining which scientific work is worthy of attention, as they are a relatively small group of people with a deciding vote in the peer review process of a large number of papers. They are often referred to as gatekeepers for this reason (Crane 1967). Our proposal entails significantly decreasing both the prevalence and importance of this role. By replacing some of this importance with long-run credit, which comes from the scientific community as a whole, it makes the evaluation of scientific work a more democratic process. Not only is there some reason to think that democratic evaluation of scientific claims is more in line with general communal norms accepted within science (Bright et al. 2018), but general arguments from democratic theory and social epistemology of science give epistemic reason to welcome the increased independence of judgment and evaluation this would introduce (List and Goodin 2001, Heesen et al. forthcoming, Perović et al. 2016, 103–104).

4 Where Peer Review Makes No Difference

In this section we consider a number of aspects of the scientific incentive structure for which we think a case can be made that abolishing peer review will leave them basically unaffected. This serves partially to forestall objections to our proposal that we anticipate from defenders of the peer review system, and partially to avoid overstating our case—in some of what follows we argue that abolishing peer review will likely have no effect in cases where one might have expected it to be beneficial.

4.1 Epistemic Sorting

Given the stated purpose of peer review mentioned in section 2 the first and most apparent disadvantage of our proposal is that it would remove the epistemic filter on what enters into the scientific literature. One might worry that the scientific community would lose the ability to maintain its own epistemic standards, and thus the general quality of scientific research would be reduced. We argue here that despite the intuitive support this idea might have, the present state of the literature on scientific peer review does not support it.

Separate out two kind of epistemic standards one may hope that the peer review system maintains. First, that peer review allows us to identify especially meritorious work and place it in high profile journals, while ensuring that especially shoddy work is kept from being published. Call this the 'epistemic sorting' function of peer review. Second, that peer review allows for the early detection of fraudulent work or work that otherwise involves research misconduct. Call this the 'malpractice detection' function of peer review. We deal with each of these in turn.

Let us step back and ask why, from the point of view of epistemic consequentialism, one would want peer review to do any sort of epistemic sorting. We take the answer to be that epistemic sorting helps scientists fruitfully direct their time and energy by selecting the best work and bringing it to scientists' attention through publication in journals. They read and respond to that which is most likely to help them advance knowledge in their field.

How could peer review achieve this? One might hope that peer review functions by keeping bad manuscripts out of the published literature and letting good manuscripts in. This, however, is a non-starter. There are far too many journals publishing far too many things, with standards of publication varying far too wildly between them, for the sheer fact of having passed peer review somewhere to be all that informative as to the quality of a manuscript.

Instead, if peer review is to serve anything like this purpose it must be because reviewers are able (even if imperfectly) to discern the relative degree of scientific merit of a work, and sort it into an appropriately prestigious journal. Epistemic sorting happens not via the binary act of granting or withholding publication, but rather through sorting manuscripts into journals located on a prestige hierarchy that tracks scientific merit.

A necessary condition for epistemic sorting to work as advertised is that reviewers be reliable guides to the merit of the scientific work they review. Our first critique is that this necessary condition does not seem to be met. Investigation into reviewing practices has not generally found much inter-reviewer reliability in their evaluations (Peters and Ceci 1982, Ernst et al. 1993, Lee et al. 2013, 5–6). What this means is that one generally cannot predict what one reviewer will think of a manuscript by seeing what another reviewer thought. If there was some underlying epistemic merit scientists were accurately (even if falteringly) discerning by means of their reviews, one would expect there to be correlations in reviewers evaluations. However, this is not what we find. Indeed, one study of a top medical journal even found that "reviewers. . . agreed on the disposition of manuscripts at a rate barely exceeding what would be expected by chance" (Kravitz et al. 2010, 3). Findings like these are typical in the literature that looks at inter-reviewer reliability (for a review of the literature see Bornmann 2011, 207). The available evidence does not provide much support for the idea that pre-publication peer review detects the presence of some underlying quality.

Our second critique of the epistemic sorting idea speaks more directly to the ideal it tracks. We are not persuaded that the best way to direct scientists' attention is to continually alert them to the best pieces of individual work, and have them proportion their attention according to position on a prestige hierarchy. We take it the intuition behind this is a broadly meritocratic one. This intuition has been challenged by some modeling work (Zollman 2009). While Zollman retained some role for peer review, his model

still found that striving to select the best work for publication is not necessarily best from the perspective of an epistemic community; his model favored a greater degree of randomization.

We do not wish to rest our case on the results of one model which in any case does not fully align with our argument, but it highlights that the ideal of meritocracy stands in need of more defense than it is typically given. We take it that scientists most fruitfully direct their attention to that package of previous work and results which, when combined, provides them with the sort of information and perspectives they need to best advance their own epistemically valuable projects. It is a presently undefended assumption that this package of work should be composed of works which are themselves individually the most meritorious work, or that paying attention to the prestige hierarchy of journals and proportioning one's attention accordingly will be useful in constructing such a package. Hence, even if it did turn out that the peer review system could sort according to scientific merit, it is an underappreciated but important fact that this is not the end of the argument. Further defense of the purpose of this kind of epistemic sorting is needed from the point of view of epistemic consequentialism.

Before moving on we note a potential objection. Even if one did not think that peer review was detecting some underlying quality or interestingness, one might think that the process of feedback and revision which forms part of the peer review system would be beneficial to the epistemic quality of the scientific literature. In this way epistemic sorting may have a positive epistemic effect even if it fails in its primary task.

However, this returns us to the points regarding gatekeepers and time allocation from section 3. We are not opposed to scientists reading each other's work, offering feedback, and updating their work in light of that. This can indeed lead to improvements (Bornmann 2011, 203), though in this context it is worth noting the results of an experiment in the biomedical sciences, which found that attempting to attach the allure of greater prestige

to more epistemically high caliber work did little to actually improve the quality of published literature (Lee 2013). Fully interpreting these results would require discussion of the measures of quality used in such literature. We do not intend to do that here, since we do not intend to dispute the point that it is desirable for scientists to give feedback and respond to it.

We would expect this sort of peer-to-peer feedback to continue under a system without pre-publication peer review. Curiosity, informal networking, collegial responsibilities, and the credit incentives to engage with others' work and make use of new knowledge before others do; these would all be retained even without pre-publication peer review. What would be eliminated is the assignment of reviewing duties to papers that scientists did not independently decide were worth their time and attention, and the necessity of giving uptake to criticism (in order to publish) independently of an author's own assessment of the value of that feedback.

We thus conclude that, from the point of view of epistemic consequentialism, there is presently little reason to believe that a loss of the epistemic sorting function of pre-publication peer review would be a loss to science. Inclusion in the literature does not do much to vouch for the quality of a paper; the evidence does not favor the hypothesis that reviewers are selecting for some latent epistemic quality in order to sort into appropriate journals; and the ideal underlying the claimed benefits of epistemic sorting is dubious. While peer reviewers do give potentially valuable feedback, there is no particular reason to think that changes in how scientists decide to spend their time would make things worse in this regard, and (per our arguments in section 3) some reason to think that they would make things better.

4.2 Malpractice Detection

The other way peer review might uphold epistemic standards is through malpractice detection. However, once again, the literature does not support this. A number of prominent cases of fraudulent research managed to sail

through peer review. Upon investigation into the behavior of those involved it was found there was no reason to think that peer reviewers or editors were especially negligent in their duties (Grant 2002, 3). Peer reviewers report unwillingness to challenge something as fraudulent even where they have some suspicion that this is so, and avoid the charge (Francis 1989, 11–12). A criminologist who looked into fraudulent behavior in science reported that "virtually no fraudulent procedures have been detected by referees because reading a paper is neither a replication nor a lie-detecting device" (Ben-Yehuda 1986, 6). A more recent survey of the evidence found, at the least, no consistent pattern in journals' self-reported ability to detect and weed out fraudulent results (Anderson et al. 2013, 235).

Even if the prospect of peer review puts some people off committing fraud, the fact that it is so unreliable at detecting fraud suggests that this is a very fragile deterrence system indeed. Even this psychological deterrence would be rapidly undermined by more adventurous souls, or those pushed by desperation, since many would quickly learn that pre-publication peer review is a paper tiger.

Conversely, there are various ways for malpractice detection to operate in the absence of peer review. These include motive modification (Nosek et al. 2012, Bright 2017a), encouraging post-publication replication and scrutiny (Bruner 2013, Romero 2017), and the sterner inculcation of the norms of science coupled with greater expectation of oversight among coworkers (Braxton 1990). All of these methods of deterring fraud or meliorating its effects would still be available under our proposal.

What evidence we now have gives little reason to suppose that abolishing pre-publication peer review is any great loss to malpractice detection. Thus in this regard our proposal would make no great difference to the epistemic health of science. Combining this with the discussion of epistemic sorting, we conclude there is presently no reason to believe pre-publication peer review is adding much value to science by upholding epistemic standards.

4.3 Herding Behavior

Where above we argued that pre-publication peer review is not making a positive difference often claimed for it, in this section we downplay a potential benefit of our proposal. A consistent worry about scientific behavior is that it is subject to fads or, in any case, some sort of undesirable herding behavior (see, e.g. Chargaff 1976, Abrahamson 2009, Strevens 2013). A natural thought is that pre-publication peer review encourages this, since by its nature it means that to get new ideas out there one must convince one's peers that the work is impressive and interesting. It has thus been claimed that pre-publication peer review encourages unambitious within-paradigm work that unduly limits the range of scientific activity (Francis 1989, 12). Reducing the incentive to herd might thus be claimed as a potential benefit of our proposal. However, we are not convinced that it is pre-publication peer review that is doing the harmful work here.

As mentioned above, our proposal eliminates or significantly reduces the importance of short-run credit, the credit that accrues to one in virtue of publishing in a (more or less prestigious) scientific journal. Long-run credit, on the other hand, is left untouched. Under any sort of credit system, a scientist needs to do work that the community will pay attention to, build upon, and recognize her for. The mere fact that (she believes that) her peers are interested in a topic and liable to respond to it is thus still positive reason to adopt a topic. This is true even if the scientist would not judge that topic to be the best use of her time if she were (hypothetically) free from the social pressures and constraints of the scientific credit system.

The best that could be said about our proposal in this regard is that scientists would not specifically have to pass a jury of peers before getting their work out there. But given that we anticipate continued competition for the attention of scientific coworkers, it is hard to say what the net effect in encouraging more experimental or less conformist scientific work would be.

Whatever conformist effects the credit incentive has (see also the discus-

sion immediately below) do not depend on whether it is short- or long-run credit one seeks. The conformism comes from the fact that credit incentives focus scientists' attention on the predicted reaction of their fellow scientists to their work. Pre-publication peer review might make this fact especially salient by bringing manuscripts before a jury of peers before they may be entered into the literature. But even without pre-publication peer review the credit-seeking scientist must be focused on her peers' opinions. So there is no particular reason to think that removing the pre-publication scrutiny of manuscripts will free scientists from their own anticipations of the fads and fashions of their day.

4.4 Long-Run Credit

We end this section by noting that many of the effects of the credit economy of science studied by social epistemologists really concern long-run credit rather than the short-run credit affected by retaining or eliminating pre-publication peer review. This point is not restricted to herding behavior.

For instance, social epistemologists have studied both the incentive to collaborate, and various iniquities that can arise when scientists do not start with equal power when deciding who shall do what work and how they shall be credited (Harding 1995, Boyer-Kassem and Imbert 2015, Bruner and O'Connor 2017, O'Connor and Bruner forthcoming). Whether or not manuscripts would have to pass pre-publication peer review in order to enter the scientific literature, there would still be benefits in the long run to collaboration, and (alas) there would still be social inequalities that allow for iniquities to manifest in the scientific prestige hierarchy.

For another example, social epistemologists have studied the ways in which the credit incentive encourages different strategies for developing a research profile or molding one's scientific personality to be more or less risk-taking (Weisberg and Muldoon 2009, Alexander et al. 2015, Thoma 2015). Once again, pre-publication peer review plays no particular role in the analy-

sis. The incentives to differentiate oneself from one's peers (without straying too far from the beaten path) and to mold one's personality accordingly exist independently of pre-publication peer review.

Two especially influential streams of work in the social epistemology of science have been the study of the division of cognitive labor (Kitcher 1990, Strevens 2003), and the role of credit in providing a spur to work in situations with a risk of under-production (Dasgupta and David 1994, Stephan 1996). These two streams have directed the focus of the field, and have formed some of the chief defenses of the credit economy of science as it now stands (but see Zollman 2018, for a more critical take).

We mention them here because pre-publication peer review or short-run credit again plays no particular role in the analyses offered by these papers. What drives their results is scientists' expectation that genuine scientific achievement will be recognized with credit. As we have argued above, it is long-run credit that best tracks genuine scientific achievement, and so it is long-run rather than short-run credit that grounds scientists' expectation in this regard. So in social epistemologists' most prominent defenses of the credit economy of science, long-run credit (while not named such) is the mechanism underlying the claimed epistemic benefits of the credit economy.

5 Difficulties For Our Proposal

We have discussed some benefits that would predictably accrue from abolishing peer review and some ways in which its apparent benefits are either under-evidenced or better attributed to the effects of long-run credit, which our proposal leaves untouched. We now discuss some cases which we take to be more problematic for our proposal—but by this point we hope to have at least convinced the reader that pre-publication peer review rests on shakier theoretical grounds than its widespread acceptance may lead one to suppose.

5.1 A Guarantee For Outsiders

One purpose pre-publication peer review serves is providing a guarantee to interested but non-expert parties. Science journalists, policy makers, scientists from outside the field the manuscript is aimed at, or interested non-scientists can take the fact that something has passed peer review as a stamp of approval from the field. At a minimum, peer review guarantees that outsiders are focusing on work that has convinced at least one relatively disinterested expert that the manuscript is worthy of public viewing. Given that there are real dangers to irresponsible science journalism or public action that is seen to be based on science that is not itself trustworthy (Bright 2018, §4), and that it is hard for non-experts to make the relevant judgment calls themselves, having a social mechanism to provide this kind of guarantee for outsiders is useful.

It is difficult to predict in advance what norms would come to exist for science journalists in the absence of pre-publication peer review. We thus first and foremost call for empirical research on this issue, possibly by studying what has happened in parts of mathematics and physics that already operate broadly along the lines we suggest (Gowers 2017).

However, against the presumption that things would be worse, we have two points to make. As the recent replication crisis has made clear, the value of peer review as a stamp of approval should not be overstated. There are reasons to doubt that peer review reliably succeeds in filtering out false results. We give three of them. First, peer reviewers face difficulties in actually assessing manuscripts—and just about anything can pass peer review eventually—as discussed under the heading of 'epistemic sorting' in section 4.1. Second, there are problems with the standards we presently use to evaluate manuscripts, in particular with the infamous threshold for statistical significance used in many fields (Ioannidis 2005, Benjamin et al. 2018). And third, deeper features of the incentive structure of science make replicability problems endemic (Smaldino and McElreath 2016, Heesen 2017c). Using

peer review as a stamp of approval may just be generating expert overconfidence (Angner 2006), without the epistemic benefits of greater reliability that would back this confidence up.

For the second part of our reply, recall that it is only pre-publication peer review that we seek to eliminate. We do not object to post-publication peer review resulting in papers being selected for inclusion in journals which mark the community's approval of such work, ideally after due and broad-based evaluation. If some such system were implemented then outsiders could use inclusion in such a journal as their marker of whether work is soundly grounded in the relevant science.

If such a stamp of approval from a journal or other communally recognized institution only comes a number of months or years after something is first published then we would expect it to represent a more well-considered judgment. Note that this would not necessarily slow the diffusion of knowledge as under the present system the same paper would have spent time hidden from view going through pre-publication peer review. The end result might not even be all that different from what happens in the present system, except that post-publication peer review would take into account more of the response or uptake from the wider scientific community. Thus it would more closely approximate the considered judgment of the community, as ultimately reflected in the long-run credit accorded to the paper.

5.2 A Runaway Matthew Effect

The second problem we are less confident we can deal with is that of exacerbating the Matthew effect. This is the phenomenon, first identified by Merton (1968), of antecedently more famous authors being credited more for work done simultaneously or collaboratively, even if the relative size or skill of their contribution does not warrant a larger share of the reward. Arguably the present system helps put a damper on the Matthew effect, allowing a junior or less prestigious author to secure attention for their work by publishing

in a high profile journal. Without such a mechanism to grab the attention of the field, perhaps scientists would just decide what to pay attention to based on their prior knowledge of the author or recommendation from others. This would strengthen the effects of networks of patronage and prestige bias favoring fancy universities. Thus squandering valuable opportunities to learn from those who were not initially lucky in securing a prestigious position or patronage from the already established.

While some have defended the Matthew effect (Strevens 2006), we will not go that route in defending our proposal for two reasons. First, the Matthew effect can perpetuate iniquities that themselves harm the generation and dissemination of knowledge (Bruner and O'Connor 2017). Second, even if it could be justified at the level of individual publications, its long-term effects are epistemically harmful. The scientific community allocates the resources necessary for future work on the basis of its recognition of past performance. So if there is excess reward for some and unfair passing over of others at the present stage of inquiry, this will ramify through to future rounds of inquiry, misallocating resources to people whose accomplishments do not fully justify their renown (Heesen 2017a). Hence on grounds of epistemic consequentialism we take seriously the problem of a runaway Matthew effect.

As mentioned, due to the pressures of credit-seeking and their own curiosity, scientists would still have incentive to read others' work and adapt it to suit their own projects. There is always a chance that valuable knowledge may be gathered from the work of one who has been ignored, which could provide an innovative edge. To some extent this creates opportunities for arbitrage: if the Matthew effect ever became especially severe there would be a credit incentive to specialize in seeking out the work of scientists who are not getting much attention. The lesson here is that the Matthew effect can only ever be so severe, before the credit incentive starts providing counter-veiling motivations.

However, this does not fully solve our problem. Moreover, so long as

resource allocation is tied to recognition of past performance the differences in recognition generated by the Matthew effect can and often do become self-fulfilling prophesies, as those with more gain the resources to do better in the future, and those without are starved of the resources necessary to show their worth.

It is not clear where to go from here. From the above it may seem like a solution would be to pair our proposal with a call to loosen the connection between recognition of a scientist's greatness based on their past performance and resource allocation. Indeed, this may well be independently motivated (Avin forthcoming, Heesen 2017a, §6). However, even short of this far-reaching change, we feel at present that this matter deserves more study rather than any definitive course of action.

Our present thought is that this is a very speculative objection, and there is no empirical evidence to back up the claim that eliminating pre-publication peer review will have dire consequences in this regard. In particular, while the present system may (rarely) allow a relative outsider to make a big splash, the common accusation of prestige bias in peer review (Lee et al. 2013, 7) suggests that on the whole pre-publication peer review may contribute to the Matthew effect rather than curtailing it.

More specifically, the Matthew effect can be made worse by peer review when anonymity breaks down in ways that systematically favor antecedently famous scientists. If this gives famous scientists more opportunities to publish papers, then our system may provide welcome relief, since it allows more people to get their papers out there. Hence whether our proposal makes the Matthew effect worse or better depends on whether the stronger influence would be who gets into the conversation (for which pre-publication peer review can exacerbate the Matthew effect), or who gets listened to once the conversation has begun (for which our proposal looks more problematic). Presently we cannot say which is the more significant effect. So, while we grant that a runaway Matthew effect may occur under our system, we prefer

to stress that at this point it is just not known whether the Matthew effect will be worse with or without pre-publication peer review.

What we propose is a large change, involving freeing up a lot of time and opening it up to more self-direction on the part of scientists, and it is not clear what sort of institutional changes it would be paired with. With more study of epistemic mechanisms designed especially to promote the work of junior or less prestigious scientists there might be found some way of surmounting the problem of a runaway Matthew effect, should it arise. Ultimately, only empirical evidence can settle these questions. Given the clear benefits and the unclear downsides of our proposal, we hope at minimum to inspire a more experimental attitude towards peer review.

6 Conclusion

Pre-publication peer review is an enormous sink of scientists' time, effort, and resources. Adopting the perspective of epistemic consequentialism and reviewing the literature on the philosophy, sociology, and social epistemology of science, we have argued that we can be confident that there would be benefits from eliminating this system, but have no strong reasons to think there will be disadvantages. There is hence a kind of weak dominance or Pareto argument in favor of our proposal.

To simplify things, imagine forming a decision matrix, with rows corresponding to 'Keeping pre-publication peer review' and 'Eliminating pre-publication peer review'. The columns would each be labeled with an issue studied by science scholars which we have surveyed here: gender bias in the literature, speed of dissemination of knowledge, efficient allocation of scientists' time and attention, etc. For each column, if there is a clear reason to think that either keeping or eliminating pre-publication scientific peer review does better according to the standards of epistemic consequentialism, place a 1 in the row of that option, and a 0 in the other. If there is no reason to

favor either according to present evidence, put a 0 in both rows.

Our present argument could then be summarized with: as it stands, the only 1s in such a table would appear in the row for eliminating pre-publication peer review. We thus advocate eliminating pre-publication peer review. Journals could still exist as a forum for recognizing and promoting work that the community as a whole perceives as especially meritorious and wishes to recommend to outsiders. Scientists would still have every reason to read, respond to, and consider the work of their peers; pre-publication peer review is not the primary drive behind either the intellect's curiosity or the will's desire for recognition, and either of those suffice to motivate such behaviors.

The overall moral to be drawn mirrors that of our invocation of the importance of long-run over short-run credit. The best guarantor of the long run epistemic health of science is science: the organic engagement with each others' ideas and work that arises from scientists deciding for themselves how to allocate their cognitive labor, and doing the hard work of replicating and considering from new angles those ideas that have been opened up to the scrutiny of the community. All this would continue without pre-publication peer review, and the best you can say for the system that currently uses up so much of our time and resources is that it often fails to get in the way.

References

Eric Abrahamson. Necessary conditions for the study of fads and fashions in science. *Scandinavian Journal of Management*, 25(2):235–239, 2009. doi: 10.1016/j.scaman.2009.03.005. URL `http://dx.doi.org/10.1016/j.scaman.2009.03.005`.

Jason McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. Epistemic landscapes, optimal search, and the division of cognitive

labor. *Philosophy of Science*, 82(3):424–453, 2015. doi: 10.1086/681766. URL http://dx.doi.org/10.1086/681766.

Melissa S. Anderson, Emily A. Ronning, Raymond De Vries, and Brian C. Martinson. Extending the Mertonian norms: Scientists' subscription to norms of research. *The Journal of Higher Education*, 81(3):366–393, 2010. ISSN 1538-4640. doi: 10.1353/jhe.0.0095. URL https://muse.jhu.edu/journals/journal_of_higher_education/v081/81.3.anderson.html.

Melissa S. Anderson, Marta A. Shaw, Nicholas H. Steneck, Erin Konkle, and Takehito Kamata. Research integrity and misconduct in the academic profession. In Michael B. Paulsen, editor, *Higher Education: Handbook of Theory and Research*, volume 28, chapter 5, pages 217–261. Springer, Dordrecht, 2013. doi: 10.1007/978-94-007-5836-0_5. URL http://dx.doi.org/10.1007/978-94-007-5836-0_5.

Erik Angner. Economists as experts: Overconfidence in theory and practice. *Journal of Economic Methodology*, 13(1):1–24, 2006. doi: 10.1080/13501780600566271. URL http://dx.doi.org/10.1080/13501780600566271.

Shahar Avin. Centralised funding and epistemic exploration. *The British Journal for the Philosophy of Science*, forthcoming. doi: 10.1093/bjps/axx059. URL http://dx.doi.org/10.1093/bjps/axx059.

Nachman Ben-Yehuda. Deviance in science: Towards the criminology of science. *British Journal of Criminology*, 26(1):1–27, 1986. doi: 10.1093/oxfordjournals.bjc.a047577. URL http://dx.doi.org/10.1093/oxfordjournals.bjc.a047577.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical

significance. *Nature Human Behaviour*, 2(1):6–10, 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0189-z. URL http://dx.doi.org/10.1038/s41562-017-0189-z.

Lutz Bornmann. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245, 2011. ISSN 1550-8382. doi: 10.1002/aris.2011.1440450112. URL http://dx.doi.org/10.1002/aris.2011.1440450112.

Thomas Boyer. Is a bird in the hand worth two in the bush? Or, whether scientists should publish intermediate results. *Synthese*, 191(1): 17–35, 2014. ISSN 0039-7857. doi: 10.1007/s11229-012-0242-4. URL http://dx.doi.org/10.1007/s11229-012-0242-4.

Thomas Boyer-Kassem and Cyrille Imbert. Scientific collaboration: Do two heads need to be more than twice better than one? *Philosophy of Science*, 82(4):667–688, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/682940.

John M. Braxton. Deviance from the norms of science: A test of control theory. *Research in Higher Education*, 31(5):461–476, 1990. doi: 10.1007/BF00992713. URL http://dx.doi.org/10.1007/BF00992713.

Liam Kofi Bright. On fraud. *Philosophical Studies*, 174(2):291–310, 2017a. ISSN 1573-0883. doi: 10.1007/s11098-016-0682-7. URL http://dx.doi.org/10.1007/s11098-016-0682-7.

Liam Kofi Bright. Decision theoretic model of the productivity gap. *Erkenntnis*, 82(2):421–442, 2017b. ISSN 1572-8420. doi: 10.1007/s10670-016-9826-6. URL http://dx.doi.org/10.1007/s10670-016-9826-6.

Liam Kofi Bright. Du Bois' democratic defence of the value free ideal. *Synthese*, 195(5):2227–2245, 2018. ISSN 1573-0964.

doi: 10.1007/s11229-017-1333-z. URL http://dx.doi.org/10.1007/s11229-017-1333-z.

Liam Kofi Bright, Haixin Dang, and Remco Heesen. A role for judgment aggregation in coauthoring scientific papers. *Erkenntnis*, 83(2):231–252, 2018. ISSN 1572-8420. doi: 10.1007/s10670-017-9887-1. URL http://dx.doi.org/10.1007/s10670-017-9887-1.

Justin Bruner and Cailin O'Connor. Power, bargaining, and collaboration. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 7, pages 135–157. Oxford University Press, Oxford, 2017.

Justin P. Bruner. Policing epistemic communities. *Episteme*, 10(4):403–416, Dec 2013. ISSN 1750-0117. doi: 10.1017/epi.2013.34. URL http://dx.doi.org/10.1017/epi.2013.34.

Erwin Chargaff. Triviality in science: A brief meditation on fashions. *Perspectives in Biology and Medicine*, 19(3):324–333, 1976. doi: 10.1353/pbm.1976.0011. URL http://dx.doi.org/10.1353/pbm.1976.0011.

Diana Crane. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 2(4):195–201, 1967. ISSN 00031232. URL http://www.jstor.org/stable/27701277.

Partha Dasgupta and Paul A. David. Toward a new economics of science. *Research Policy*, 23(5):487–521, 1994. ISSN 0048-7333. doi: 10.1016/0048-7333(94)01002-1. URL http://www.sciencedirect.com/science/article/pii/0048733394010021.

Margaret Eisenhart. The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, 32(2):241–255, 2002. ISSN 1573-1898. doi: 10.1023/A:1016082229411. URL http://dx.doi.org/10.1023/A:1016082229411.

Edzard Ernst, T. Saradeth, and Karl Ludwig Resch. Drawbacks of peer review. *Nature*, 363(6427):296, 1993. doi: 10.1038/363296a0. URL `http://dx.doi.org/10.1038/363296a0`.

Henry Etzkowitz, Stefan Fuchs, Namrata Gupta, Carol Kemelgor, and Marina Ranga. The coming gender revolution in science. In Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, editors, *The Handbook of Science and Technology Studies*, chapter 17, pages 403–428. MIT Press, Cambridge, third edition, 2008. ISBN 9780262083645.

Daniele Fanelli. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, 5(4):e10271, Apr 2010. doi: 10.1371/journal.pone.0010271. URL `http://dx.doi.org/10.1371/journal.pone.0010271`.

Jere R. Francis. The credibility and legitimation of science: A loss of faith in the scientific narrative. *Accountability in Research: Policies and Quality Assurance*, 1(1):5–22, 1989. doi: 10.1080/08989628908573770. URL `http://dx.doi.org/10.1080/08989628908573770`.

Alvin I. Goldman. *Knowledge in a Social World*. Oxford University Press, Oxford, 1999. ISBN 0198237774.

Timothy Gowers. The end of an error? *The Times Literary Supplement*, October 2017. URL `https://www.the-tls.co.uk/articles/public/the-end-of-an-error-peer-review/`. Editorial.

Paul M. Grant. Scientific credit and credibility. *Nature Materials*, 1:139–141, 2002. doi: 10.1038/nmat756. URL `http://dx.doi.org/10.1038/nmat756`.

Sandra Harding. "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3):331–349, 1995. doi: 10.1007/BF01064504. URL `http://dx.doi.org/10.1007/BF01064504`.

Remco Heesen. Academic superstars: Competent or lucky? *Synthese*, 194 (11):4499–4518, 2017a. ISSN 1573-0964. doi: 10.1007/s11229-016-1146-5. URL http://dx.doi.org/10.1007/s11229-016-1146-5.

Remco Heesen. Communism and the incentive to share in science. *Philosophy of Science*, 84(4):698–716, 2017b. ISSN 0031-8248. doi: 10.1086/693875. URL http://dx.doi.org/10.1086/693875.

Remco Heesen. Why the reward structure of science makes reproducibility problems inevitable. Manuscript, September 2017c. URL http://remcoheesen.files.wordpress.com/2015/03/rewards-and-reproducibility2.pdf.

Remco Heesen. When journal editors play favorites. *Philosophical Studies*, 175(4):831–858, 2018. ISSN 0031-8116. doi: 10.1007/s11098-017-0895-4. URL http://dx.doi.org/10.1007/s11098-017-0895-4.

Remco Heesen, Liam Kofi Bright, and Andrew Zucker. Vindicating methodological triangulation. *Synthese*, forthcoming. ISSN 1573-0964. doi: 10.1007/s11229-016-1294-7. URL http://dx.doi.org/10.1007/s11229-016-1294-7.

Erin Hengel. Publishing while female: Are women held to higher standards? Evidence from peer review. Manuscript, August 2018. URL http://www.erinhengel.com/research/publishing_female.pdf.

David L. Hull. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. The University of Chicago Press, Chicago, 1988. ISBN 0226360504.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, Aug 2005. doi: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

Saana Jukola. A social epistemological inquiry into biases in journal peer review. *Perspectives on Science*, 25(1):124–148, 2017. doi: 10.1162/POSC_a_00237. URL http://dx.doi.org/10.1162/POSC_a_00237.

J. Katzav and K. Vaesen. Pluralism and peer review in philosophy. *Philosophers' Imprint*, 17(19):1–20, 2017. URL http://hdl.handle.net/2027/spo.3521354.0017.019.

Philip Kitcher. The division of cognitive labor. *The Journal of Philosophy*, 87(1):5–22, 1990. ISSN 0022362X. URL http://www.jstor.org/stable/2026796.

Richard L. Kravitz, Peter Franks, Mitchell D. Feldman, Martha Gerrity, Cindy Byrne, and William M. Tierney. Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? *PLoS ONE*, 5(4):e10072, 2010. doi: 10.1371/journal.pone.0010072. URL http://dx.doi.org/10.1371/journal.pone.0010072.

Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, second edition, 1986.

Carole J. Lee. The limited effectiveness of prestige as an intervention on the health of medical journal publications. *Episteme*, 10(4):387–402, 2013. doi: 10.1017/epi.2013.35. URL http://dx.doi.org/10.1017/epi.2013.35.

Carole J. Lee. Revisiting current causes of women's underrepresentation in science. In Jennifer Saul and Michael Brownstein, editors, *Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology*, chapter 2.5, pages 265–282. Oxford University Press, Oxford, 2016. doi: 10.1093/acprof:oso/9780198713241.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780198713241.001.0001.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

Christin List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001. ISSN 1467-9760. doi: 10.1111/1467-9760.00128. URL http://dx.doi.org/10.1111/1467-9760.00128.

Helen E. Longino. *Science as Social Knowledge*. Princeton University Press, 1990.

Karen Seashore Louis, Lisa M. Jones, and Eric G. Campbell. Macroscope: Sharing in science. *American Scientist*, 90(4):304–307, 2002. ISSN 00030996. URL http://www.jstor.org/stable/27857685.

Bruce Macfarlane and Ming Cheng. Communism, universalism and disinterestedness: Re-examining contemporary support among academics for Merton's scientific norms. *Journal of Academic Ethics*, 6(1):67–78, 2008. ISSN 1570-1727. doi: 10.1007/s10805-008-9055-y. URL http://dx.doi.org/10.1007/s10805-008-9055-y.

Robert K. Merton. A note on science and democracy. *Journal of Legal and Political Sociology*, 1(1–2):115–126, 1942. Reprinted in Merton (1973, chapter 13).

Robert K. Merton. Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22(6):635–659, 1957. ISSN 00031224. URL http://www.jstor.org/stable/2089193. Reprinted in Merton (1973, chapter 14).

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63,

1968. ISSN 00368075. URL http://www.jstor.org/stable/1723414. Reprinted in Merton (1973, chapter 20).

Robert K. Merton. Behavior patterns of scientists. *The American Scholar*, 38 (2):197–225, 1969. ISSN 00030937. URL http://www.jstor.org/stable/41209646. Reprinted in Merton (1973, chapter 15).

Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago, 1973. ISBN 0226520919.

Brian A. Nosek, Jeffrey R. Spies, and Matt Motyl. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012. doi: 10.1177/1745691612459058. URL http://pps.sagepub.com/cgi/content/abstract/7/6/615.

Cailin O'Connor and Justin Bruner. Dynamics and diversity in epistemic communities. *Erkenntnis*, forthcoming. ISSN 1572-8420. doi: 10.1007/s10670-017-9950-y. URL http://dx.doi.org/10.1007/s10670-017-9950-y.

Slobodan Perović, Sandro Radovanović, Vlasta Sikimić, and Andrea Berber. Optimal research team composition: data envelopment analysis of Fermilab experiments. *Scientometrics*, 108(1):83–111, 2016. doi: 10.1007/s11192-016-1947-9. URL http://dx.doi.org/10.1007/s11192-016-1947-9.

Douglas P. Peters and Stephen J. Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2):187–195, 1982. doi: 10.1017/S0140525X00011213. URL http://dx.doi.org/10.1017/S0140525X00011213.

Katarina Prpić. Gender and productivity differentials in science. *Scientometrics*, 55(1):27–58, 2002. ISSN 0138-9130. doi: 10.1023/A:1016046819457. URL http://dx.doi.org/10.1023/A:1016046819457.

RIN. Activities, costs and funding flows in the scholarly communications system in the UK. Technical report, Cambridge Economic Policy Associates on behalf of the Research Information Network, 2008. URL http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/activities-costs-and-funding-flows-scholarly-commu.

Felipe Romero. Novelty versus replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, 84(5):1031–1043, 2017. ISSN 0031-8248. doi: 10.1086/694005. URL http://dx.doi.org/10.1086/694005.

Jennifer Saul. Implicit bias, stereotype threat, and women in philosophy. In Katrina Hutchison and Fiona Jenkins, editors, *Women in Philosophy: What Needs to Change?*, chapter 2, pages 39–60. Oxford University Press, Oxford, 2013.

Paul E. Smaldino and Richard McElreath. The natural selection of bad science. *Royal Society Open Science*, 3(9), 2016. doi: 10.1098/rsos.160384. URL http://rsos.royalsocietypublishing.org/content/3/9/160384.

Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182, 2006. URL http://jrs.sagepub.com/content/99/4/178.short.

Paula E. Stephan. The economics of science. *Journal of Economic Literature*, 34(3):1199–1235, 1996. URL http://www.jstor.org/stable/2729500.

Michael Strevens. The role of the priority rule in science. *The Journal of Philosophy*, 100(2):55–79, 2003. ISSN 0022362X. URL http://www.jstor.org/stable/3655792.

Michael Strevens. The role of the Matthew effect in science. *Studies in History and Philosophy of Science Part A*, 37(2):159–170, 2006. ISSN 0039-3681. doi: http://dx.doi.org/10.1016/j.shpsa.2005.07.009. URL http://www.sciencedirect.com/science/article/pii/S0039368106000252.

Michael Strevens. Herding and the quest for credit. *Journal of Economic Methodology*, 20(1):19–34, 2013. doi: 10.1080/1350178X.2013.774849. URL http://dx.doi.org/10.1080/1350178X.2013.774849.

Michael Strevens. Scientific sharing: Communism and the social contract. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 1. Oxford University Press, Oxford, 2017. URL https://philpapers.org/rec/STRSSC-2.

Johanna Thoma. The epistemic division of labor revisited. *Philosophy of Science*, 82(3):454–472, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/681768.

Virginia Valian. *Why So Slow? The Advancement of Women*. MIT Press, Cambridge, 1999. ISBN 9780262720311.

Richard Van Noorden. The true cost of science publishing. *Nature*, 495 (7442):426–429, 2013. ISSN 0028-0836. doi: 10.1038/495426a. URL http://dx.doi.org/10.1038/495426a.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/644786.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6(2):185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://dx.doi.org/10.3366/E174236000900063X.

Kevin J. S. Zollman. The credit economy and the economic rationality of science. *The Journal of Philosophy*, 115(1):5–33, 2018. doi: 10.5840/jphil201811511. URL http://dx.doi.org/10.5840/jphil201811511.

Harriet Zuckerman and Jonathan R. Cole. Women in American science. *Minerva*, 13(1):82–102, 1975. ISSN 1573-1871. doi: 10.1007/BF01096243. URL http://dx.doi.org/10.1007/BF01096243.

Epistemic Loops and Measurement Realism

Alistair M. C. Isaac

Abstract

Recent philosophy of measurement has emphasized the existence of both diachronic and synchronic "loops," or feedback processes, in the epistemic achievements of measurement. A widespread response has been to conclude that measurement outcomes do not convey interest-independent facts about the world, and that only a coherentist epistemology of measurement is viable. In contrast, I argue that a form of measurement realism is consistent with these results. The insight is that antecedent structure in measuring spaces constrains our empirical procedures such that successful measurement conveys a limited, but veridical knowledge of "fixed points," or stable, interest-independent features of the world.

§1 Introduction

Recent philosophy of measurement has employed detailed case studies to highlight the complex, iterative process by which measurement practices are refined. Typically, these examples are taken to support some form of epistemic coherentism, on which the validation of measurement procedures, and thus their epistemic import, is irreducibly infected by the contingent history of their development in aid of human interests. This coherentism in turn undermines *measurement realism*, the view that outcomes of successful measurement practices veridically represent objective (i.e. interest-independent) features of the world. For instance, van Fraassen (2008) takes the historical contingency of measurement practice to support empiricism, and Chang (2012) argues that only a pragmatic, interest-relative "realism" about measurement outcomes is plausible, not one which interprets them as corresponding to objective features in the world. More generally, Tal (2013) identifies coherentism as a major trend within contemporary philosophy of measurement.

I argue that the iterative and coherentist features of measurement practice these authors rightly emphasize are nevertheless consistent with realism about measurement outcomes. Nevertheless, my position contrasts significantly with that of other measurement realists, such as Byerly and Lazara (1973) or Michell (2005), who take measurement realism to be continuous with global scientific realism. On their view, measurement realism is a *stronger* position than traditional realism, imputing reality not only to theoretical objects and laws, but also to their quantitative character. The view defended here reverses this priority, articulating a realism about measurement outcomes *weaker* than traditional realism. In particular, I argue that the convergent assignment of increasingly precise values that constitutes successful measurement serves as incontrovertible evidence for *fixed points* in the world — features or events standing in stable quantitative relationships — even though the evidence it provides for any non-numerical theoretical description of these points is defeasible. The insight here is that measurement is more evidentially demanding than traditional confirmation, i.e. it requires a greater contribution from the interest-independent world to succeed than mere qualitative experiments. I argue that this greater evidential demand is a consequence of the

antecedent numerical structure in which measurement outcomes are represented. This antecedent structure blocks the possibility of gerrymandered categories that crosscut the joints of nature. Consequently, successful measurement constitutes a substantive enough epistemic achievement that we may legitimately "factor out" the contribution to success made by human interests, and accept the outcome as representing an objective feature of the world.

After surveying the motivations for measurement coherentism, I elaborate on the notion of "successful" measurement, and why it poses a challenge to coherentism. The paper concludes with a more careful articulation of the distinctive features of fixed point realism.

§2 Epistemic Loops in Measurement Practice

Contemporary measurement coherentism is motivated by two types of case study, each identifying a different kind of epistemic "loop," or feedback process driving knowledge formation. Chang and van Fraassen emphasize diachronic examples of epistemic iteration, where the feedback process extends over several stages of mutual influence between theory change and refinement of measurement practice. A different kind of epistemic loop has been discussed by Tal and metrologist Mari, who highlight the role of models in the calibration of measurement instruments and the assignment of quantity values, illustrating a synchronic epistemic interdependence between theory and measurement.

§2.1 Epistemic Iteration

Chang (2004) defines *epistemic iteration* as "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals" (45). He takes this process to support a "progressive coherentism": on the one hand, the criteria for measurement success are internal to a practice, so scientific knowledge does not rest on an independent foundation; on the other hand, these internal criteria may be used to evaluate new practices as improvements or refinements on their predecessors, thereby allowing for scientific progress (in contrast to traditional coherentism, Chang 2007). In the context of measurement, this means that later measurement practices may be understood as in some sense "better" than earlier ones, yet these "epistemic achievements" should not be cashed out as greater degree of correspondence to quantities in the world.

For instance, thermometry as a practice begins with subjective assignments of relative heat on the basis of our bodily experiences. Noticing that fluids appear to change volume in rough correspondence with these subjective sensations, one may construct a thermoscope, or device allowing comparison of relative fluid volumes in different circumstances. Already a theoretical leap is required to identify the cause of these changes in relative volume with the cause of our differing subjective sensations, especially given the discrepancies between these sensations and our thermoscopic readings (e.g. contrary to experience, caves are warmer in summer than they are in winter). Nevertheless, the move to the thermoscope constitutes an epistemic achievement, in the sense that it allows for greater regularity in the assignment of relative temperatures, both

across contexts and across observers. A similar pattern is seen in the move from thermoscope to thermometer, which enables assignment of numbers to temperatures. Numerical representation constitutes a yet greater epistemic achievement, insofar as it allows comparison of temperature assignments across devices. Nevertheless, this practice does not itself guarantee greater veracity of temperature assignments, since it rests on the assumption that temperature varies linearly with changes in the height of thermometric fluid. But this assumption cannot itself be verified, as that would require access to temperature in the world by some means independent of thermometry. Similar achievements, (seemingly) inextricably entangled with theory, may be seen at each further stage in the development of thermometric practice.

The moral of this case study is the historical contingency of thermometry, and thus of its results. At each stage in the development of thermometry, an advance in theory was required to extend measurement practice. Internal criteria of consistency and increased precision in the assignment of numerical values establish the new practice as an advance over the previous one. Yet, the application of these criteria is not empirically constrained. When one assumes that "temperature" (whatever it may be) varies linearly with changes in the height of the indicator column in an air thermometer, one is making an assumption both necessary for measurement progress and in principle non-empirical, since no independent access to "temperature," outside the behavior of the very devices and procedures under investigation, is possible: "*Prior to the construction of a thermometer, there is no thermometer to settle that question!*" (van Fraassen 2008, 126, emphasis in original). Chang (2004) argues that, in order to make sense of the "progress" exemplified by cases like these, we have to "look away from truth," and appeal only to historically contingent criteria for success (227)—"scientific progress … cannot mean closer approach to the truth" (228); "Truth, in the sense of correspondence to reality, is beyond our reach" (Chang 2007, 20). The delusion that one may evaluate the correspondence between our assignment of temperatures and the objective state of the world rests on the mistaken and "impossible god-like view in which nature and theory and measurement practice are all accessed independently of each other" (van Fraassen 2008, 139). Rather, the only relevant notion of "truth" for assessing the success of thermometry "rests first and foremost on coherence with the rest of the system" (Chang 2012, 242).

§2.2 Models and Calibration

Another kind of epistemic loop is found in synchronic measurement practice, where *models* play a constitutive role in determining measurement outcomes. The crucial concept here is *calibration*, the process of correcting a measurement device for inferred discrepancies between its readout and the target value. Calibration is a necessary feature of all sophisticated measurement, yet the process of calibration illustrates the ineliminable role of theoretical posits in the very assignment of quantity values in an act of measurement. When measuring, scientists do not (as one might naively suppose) read values directly from nature, rather they employ models of the interaction between measurement device and target system in order to "correct" the readout value to a final assigned value (Mari and Giordani 2014).

Tal (2014) illustrates this point through the example of the measurement of time, in particular coordinated universal time (UTC). The second is presently defined as 9,192,631,770 periods of the hyperfine transition between the two ground states of a caesium-133 atom at zero degrees Kelvin.[1] Models feature at every step of the process leading from devices that interact directly with caesium atoms to the UTC. First, it is impossible to probe caesium atoms at absolute zero, so the enumeration of hyperfine transitions output by a caesium clock must be corrected for this discrepancy. This, as well as other corrections, rely on models of the physical interaction between the device and the atom in order to infer the discrepancy between the actual state of the system and the idealized state referred to in the definition. Caesium clocks are too complex to run continuously, so their output is used to calibrate more mundane atomic clocks (301). Furthermore, the UTC itself is not identified with the output of any one clock; rather, it is calculated retrospectively by a weighted average over all participating atomic clocks, with weights determined by the degree of past fit between each clock and previous calculations of UTC (302–3).

The lessons of this example are analogous to those of epistemic iteration: measurement improvement appears to rest on internal standards of coherence rather than on correspondence with external quantities. The weighting procedure that leads to UTC, for instance, "promotes clocks that are stable relative to each other" (304). Success at achieving this stability indeed demonstrates "genuine empirical knowledge," but not knowledge in the first instance about a regularity in the objective world, but rather a regularity "in the behaviour of instruments" (327). Consequently, it is a "conceptual mistake" to think that "the stability of measurement standards can be analysed into distinct contributions by humans and nature" (328). On an extreme interpretation of this view, even computer simulation constitutes a form of measurement (Morrison 2009). The basic idea is that, once we grant the ineliminable role of models in measurement, it is a small conceptual step to accept that the aspect of measurement involving empirical contact with the world may be arbitrarily distant from that involving modeling (Parker 2017).

§3 Achieving Successful Measurement

For the remainder of this paper, I wish to grant the basic descriptive features of this account: both diachronically and synchronically, successful measurement involves epistemic loops. Nevertheless, I will argue, there is a form of measurement realism consistent with these loops; one on which the contingent, interest-relative, and theory-laden aspects of measurement may indeed be factored out, leaving the bare, objective facts about the world conveyed by successful measurement.

[1] Arguably, the process of establishing UTC is not measurement at all — since the length of the second is *defined* by caesium-133 transitions, it is not subject to empirical determination. The purpose of the project Tal examines is not to establish a value, as in paradigmatic cases of measurement, but rather to coordinate time-relevant activities across the globe with maximal precision. I set this concern aside for the discussion here, since Tal's analysis has been so influential in philosophy of measurement, and his conclusions concerning the model-mediation of measurement incontrovertibly reflect the practices of metrologists.

But what is "successful measurement"? For the purposes of discussion here, I take *measurement* to be any empirical procedure for assigning points (or regions) in a metric space to states of the world, where a *metric space* is any set of elements with a distance metric defined over it. This means, on the one hand, that I rule out degenerative forms of "measurement" that simply assign objects to categories, or place them in an order (the *nominal* and *ordinal* scales of Stevens 1946). On the other hand, I include measurement procedures that map states of the world into any geometrical space, not just the real line, so long as they have an assigned distance metric (siding with Suppes, et al. 1989, against Díez 1997); nevertheless, in the interests of simplicity, I will refer to these outcomes as "numerical" assignments, since they may be represented by vectors of real numbers. In line with Krantz et al. (1971), I take it that one can determine whether or not an empirical procedure constitutively requires the metric features of a geometrical space by analyzing whether these remain invariant across permissible transformations over the mapping into that space.[2]

I take *successful* measurement to exhibit two key features: *convergence* and *precision*. These features pose a significant challenge to the thoroughgoing coherentist.

§3.1 Convergence

Coherentists have emphasized the theory-ladenness of both diachronic and synchronic aspects of measurement refinement. However, a hallmark of sophisticated scientific measurement is its attempt to factor out the role of theory in measurement by employing different theoretical commitments to measure the same quantity. A measurement practice *converges* when procedures employing different theoretical commitments arrive at the same outcome.

For instance, in the early 20th century, a wide variety of phenomena were investigated, employing distinct methods and theoretical commitments, in the attempt to measure Avogadro's constant N_A, the number of particles in a mole of substance. Perhaps most well-known are Perrin's experiments on Brownian motion, which, in combination with Einstein's theoretical analysis, allowed an assignment of value to N_A. However, similar values were achieved by radically different means. For instance, Millikan was able to determine N_A by measuring charge of the electron through his oil drop experiments and dividing the Faraday constant (charge of a mole of electrons) by his result. Millikan's measurement relied on Stokes' theoretical analysis of the movement of spheres through a viscous fluid — insofar as Brownian motion was a factor, it was as a source of noise, not (as for Perrin) a source of evidence. Black body radiation and the blue

[2] For instance, consider two procedures for assigning real numbers to my students. On the first, I assign a number to each letter-type with which a student's name begins (e.g. A=1, B=3,…); on the second, I hold a meter stick up to each student and note their height. The former procedure is indifferent to the algebraic structure of the real line (letters do not add or subtract from each other systematically), and thus metric features of the real line are not invariant across alternative, equally permissible assignments of numbers (e.g. A=7, B=15,…). The second does make use of algebraic structure (as heights do "add" through concatenation), and thus metric features remain invariant across alternative assignments (Jamal is twice the height of Leslie, whether their heights are represented in inches or centimeters). So, on the present definition, the latter procedure is measurement, but the former is not.

of the sky are examples of other phenomena that, when combined with theoretical models of photon emission and diffraction respectively, allow alternate means of measuring N_A. Insofar as these procedures assign the same value to N_A, they converge.

I want to stress that the point being made here is *not* the traditional realist one, that these practices provide converging evidence for the particulate nature of matter, whether as "common cause" (Salmon 1984) or most likely hypothesis (Psillos 2011). Those arguments are instances of *abduction*, while I am interested in whether a stronger, non-abductive conclusion may be drawn from convergence. A better analogy is with the discussion of robustness in the modeling literature: a result is *robust* if it is obtained by a plurality of models that each make different simplifying assumptions (Weisberg 2006). The particulate nature of matter is not robust in this sense across different measurement practices, since it is assumed by all of them. However, the value of N_A is robust, since that value is not itself assumed, and is obtained with a great degree of agreement despite differences in the assumptions made by each measurement practice (and its supporting models). I claim that convergence toward this value provides robust, non-abductive evidence for an objective feature of the world.

This example is in no way exceptional: convergent measurement practices are rife across the sciences. Smith and Miyake, for instance, have investigated a number of examples. Thomson's convergent measurements of the charge of the electron employed a variety of different methods and assumptions (Smith 2001). Early attempts to measure the density of the interior of the earth likewise assumed a variety of different theoretical models (Miyake 2018). In more recent research, measurements of the constants that govern molecular vibration converge across spectroscopy, chemistry, thermodynamics, and femtochemistry (Smith and Miyake, *manuscript*). To pick an example from an entirely different area of science, measurements of the spectral sensitivity of mammalian retinal receptors employing psychophysical methods (extracting sensitivity curves from behavioral color matching experiments, as performed by Helmholtz in the late 19th century) converge closely with 20th century physiological methods (detecting rate of nerve firing in (e.g.) cow retinal tissue in response to single wavelength lights, Wandell 1995). In all of these cases, "What is being shown through the convergence of these measurements is that the discrepancies between the different measurements … are due to the particularities of the models being used" (Miyake, 2018, 336). In other words, convergence factors out model-sensitive features of measurement; in order for it to occur, "the empirical world has to cooperate" (Smith 2001, 26).

§3.2 Precision

Traditionally, measurement success was evaluated with respect to two features: accuracy and precision. *Accuracy* was degree of approach to true value, while *precision* was degree of specificity in the value provided. The considerations in §2 undermine the criterion of accuracy, since they show we have no independent access to "true values" and thus cannot use them as standards for evaluating measurement (Mari 2003). Nevertheless, we can still assess measurements for precision, since it may be defined operationally: a measurement is *precise* to the

extent that it returns the same result when performed repeatedly. The number of *significant figures* in a numerical assignment indicates the degree of measurement precision, since these characterize the size of the region within which repeated measurements fall.

Coherentists stress the fact that increased precision is a purely internal criterion for improving measurement. Here, however, I want to stress the way in which increased precision constitutes a qualitatively different, and more impressive, epistemic achievement than other forms of empirical success, such as qualitative prediction or improved coherence of classification. These qualitative achievements are subject to worries about semantic and theoretical holism: one may always succeed in classification, or correct qualitative prediction, by suitably redrawing the boundaries of one's theoretical concepts. As LaPorte (2004) argues, when faced with anomalies in the relationship between guinea pigs and prototypical rodents, or birds and dinosaurs, scientists face a *choice* whether to expand or contract their previous categories to include or exclude perceived outliers (a similar case is made by Slater 2017 for Pluto and planethood). Nothing about the prior conceptual framework itself forces this choice one way or another, nor do demands for internal consistency.

Measurement is different from mere categorization precisely because it maps states into a metric space. The crucial point to note here is that a metric space has *antecedent structure*: the distances between points on the real line, and the algebraic relationships between them, are fixed *before* we employ it to represent height or temperature or electric charge. This antecedent structure constrains the relationship between measurement outcomes, independently restricting our assessment of them as same or different, or converging or not, in a manner impervious to ad hoc revision. Increase in precision occurs when successive measurement practices are able to shrink distances (between repeated measurements within each practice) determined by the metric of the representing space. Thus, the metric of this space serves two functions: (i) it represents the distances between different measured quantities, but (ii) it also provides a directed metric for improving measurement of a single quantity, since it determines the distances between repeated measurements that characterizes their precision. Consequently, pace van Fraassen, attempts to increase precision are empirically constrained, since this directed metric for improvement can only be satisfied through the cooperation of nature: if nature is not sufficiently stable where we probe it, no choice, convention, or increased coherence can reduce the distances between our repeated attempts to measure it. Some examples will illustrate this point.

Consider, for instance, determinations of the boiling point of water. Chang (2004, Ch. 1) surveys the sequence of choice points in the early practice of thermometry leading to relative stability in the measurement of this temperature: what are the visual indicators of boiling, where should the thermometer be positioned, what should be the shape of the vessel holding the water, its material, etc.[3] Decisions on each of these points affect the relative stability in the thermometric reading, illustrating the naivety of a view on which

[3] The issue here is the phenomenon of "superheating," whereby water with relatively little dissolved gas, or in a flask with very small surface area, may be heated to a higher temperature without bubbling.

boiling point is a simple phenomena merely waiting to be observed. Nevertheless, in committing to represent the boiling point numerically, investigators subjected themselves to a criterion for success distinct from coherence. If the numbers assigned by thermometers within this-shaped vessels and that-shaped ones differ during phenomenologically similar bubblings, then the distance between those numbers provides a criterion of difference that must be respected if thermometric practice is to count as measurement. Restricting attention to those vessels that minimize distances between numerical outcomes is thus not a mere choice, or gerrymandering of the category "boiling," since it is forced upon the investigator by an antecedent metric for success.

Likewise, consider again the determination of UTC through the retrospective weighting of the comparison set of atomic clocks. For Tal, the success of this procedure is evidence for stability in our clocks, but not for any human-independent feature of the world. Nevertheless, UTC is constrained by the world in two distinct ways. First, through empirical contact with caesium atoms. While this contact is mediated by models, these models themselves are the result of convergent measurements of atomic phenomena through a wide variety of means, employing distinct theoretical assumptions. Second, the distance metric of the real line constrains the assessment of fit between clocks in the set. While the algorithm that weights them takes degree of internal agreement as the standard for higher weighting, the metrical structure of the space in which relative rates of the clocks are assessed ensures relative agreement cannot be stipulated, fudged, or gerrymandered. The clocks need to cooperate by performing stably enough that they may be compared with a high degree of precision, and this stable point remains tethered to a robust regularity in the world through checks with the convergent behavior of caesium.

While UTC is in some respects atypical (see footnote 1), these three features — internal coordination of outcomes, empirical checks, and directed improvement constrained by the real line — are features of scientific measurement in general. What Tal's discussion of the UTC obscures is the sheer number of empirical checks typically involved, and the strictness of the demands placed by conformity to the metric of improvement the measuring space provides. In official determinations of fundamental physical constants, convergence is demanded across *all* measurement procedures, as assessed by the law-governed interrelationship between physical quantities, and the degree of precision achieved illustrates the strictness of this demand. For instance, in late 19th century measurements of N_A by Perrin and e (charge of electron) by Thomson, only 2 to 3 significant figures were typically obtained within method, and convergence across methods often only agreed as to order of magnitude. By 1911, Millikan was measuring both e and N_A to 4 significant figures, and demonstrating that the models employed to calibrate the oil drop method converged closely with other aspects of physical theory (1911). As of 2014, N_A was being measured at upwards of 9 significant figures, and e upwards of 11 (Mohr et al. 2016).[4] In each case, the increase in precision has been constrained by the antecedent structure of the real line, and thus is not itself a matter of mere convention or coherence. Rather, the world must cooperate by remaining

[4] It is expected that after the 2018 26th General Conference on Weights and Measures, N_A and e will be fixed as constants to which other quantities may be referred during measurement.

sufficiently stable if such precision is to be possible; consequently, precise values constitute robust evidence for points of objective fixity in the world revealed through measurement.

§4 Conclusion: Fixed-Point Realism

Traditional scientific realism rests on an abductive inference from observed empirical success to presumed underlying causes. Successful measurement may certainly be used in such an inference, but I claim here that it non-abductively supports a more modest realism:

> **Fixed Point Realism** – values obtained through successful measurement veridically represent objective fixed points in the world, which may be exhaustively characterized by the pattern of distances that obtain between them in a metric space.

FPR is a form of *epistemic structural realism*. It differs from traditional realism insofar as it claims a veridical characterization of the world is possible independent of any particular theoretical description. Our theory of the nature of temperature or of state changes may change radically, yet the points of relative stability characterizing, e.g., boiling point of water, "absolute zero," freezing point of oxygen, etc., will stay robust across any such change, and that robustness may be represented by their relative positions within a numerical scale.

FPR differs from other flavors of structural realism in the type of structure to which it is committed. Structural realists typically focus on the rich mathematical structure of physical theory, and derivation or limit relations that hold between successive theories, e.g. Newton's laws are a limit case of relativistic mechanics (Worrall 1989). FPR commits itself only to *geometric* structure, i.e. the pattern of relative distances that obtain between points of stability as represented in a metric space. Just as our theoretical description of these stable points may change, so may our mathematical account of their relationship — if new mathematical physics fails to derive old equations as limit cases, this in no way jeopardizes the veridicality of this geometric structure.

Finally, FPR disagrees with coherentism, insofar as it asserts that the geometrical structure uncovered through acts of successive measurement obtains in the world independent of our practices. It does not deny the importance of epistemic loops for understanding the process of measurement. Nevertheless, it takes convergence in measured values to indicate that the points of stability they represent obtain independent of the theoretical commitments encapsulated in the models used for calibration. Likewise, it takes increased precision to constitute a criterion for measurement success over and above that of coherence, one that is only realized when the interest-independent world cooperates with us by remaining stable when we probe it.

Bibliography

Byerly, H., and V. Lazara (1973) "Realist Foundations of Measurement," *Philosophy of Science* 40:10–28.

Chang, H. (2004) *Inventing Temperature*, Oxford UP.

Chang, H. (2007) "Scientific Progress: Beyond Foundationalism and Coherentism," O'Hear (ed.) *Royal Institute of Philosophy Supplement* 61:1–20.

Chang, H. (2012) *Is Water H_2O?* Springer.

Díez, J. (1997) "A Hundred Years of Numbers: An Historical Introduction to Measurement Theory 1887–1990, part ii," *Studies in History and Philosophy of Science* 28:237–265.

Krantz, D., R. Luce, P. Suppes, and A. Tversky (1971) *Foundations of Measurement*, vol. 1, Dover.

LaPorte, J. (2004) *Natural Kinds and Conceptual Change*, Cambridge UP.

Mari, L. (2003) "Epistemology of Measurement," *Measurement* 34:17–30.

Mari, L., and A. Giordani (2014) "Modeling Measurement: Error and Uncertainty," in Boumans, Hon, and Petersen (eds.) *Error and Uncertainty in Scientific Practice*, Pickering & Chatto: 79–96.

Michell, J. (2005) "The Logic of Measurement: A Realist Overview," *Measurement* 38:285–294.

Millikan, R. (1911) "On the Elementary Electrical Charge and the Avogadro Constant," *Physical Review* 2:349–397.

Miyake, T. (2018) "Scientific Realism and the Earth Sciences," in Saatsi (ed.) *The Routledge Handbook of Scientific Realism*, Routledge: 333–344.

Mohr, P., D. Newell, and B. Taylor (2016) "CODATA Recommended Values of the Fundamental Physical Constants: 2014," *Review of Modern Physics* 88:035009.

Morrison, M. (2009) "Models, Measurement and Computer Simulation: The Changing Face of Experimentation," *Philosophical Studies* 143:33–57.

Parker, W. (2017) "Computer Simulation, Measurement, and Data Assimilation," *British Journal for Philosophy of Science* 68:273–304.

Psillos, S. (2011) "Moving Molecules above the Scientific Horizon: On Perrin's Case for Realism," *Journal for General Philosophy of Science* 42:339–363.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton UP.

Slater, M. (2017) "Plato and the Platypus: An Odd Ball and an Odd Duck – On Classificatory Norms," *Studies in History and Philosophy of Science* 61:1–10.

Smith, G. (2001) "J.J. Thomson and the Electron, 1897–1899," in Buchwald and Warwick (eds.) *Histories of the Electron*, MIT Press.

Smith, G., and T. Miyake (*manuscript*) "Realism, Physical Meaningfulness, and Molecular Spectroscopy"

Stevens, S. (1946) "On the Theory of Scales of Measurement," *Science* 103(2684):677–680.

Suppes, P., D. Krantz, R. Luce, and A. Tversky (1989) *Foundations of Measurement*, vol. 2, Dover.

Tal, E. (2013) "Old and New Problems in Philosophy of Measurement," *Philosophy Compass* 8/12:1159–1173.

Tal, E. (2014) "Making Time: A Study in the Epistemology of Measurement," *British Journal for Philosophy of Science* 67:297–335.

Wandell, B. (1995) *Foundations of Vision*, Sinauer.

Weisberg, M. (2006) "Robustness Analysis," *Philosophy of Science* 73:730–742.

Worrall, J. (1989) "Structural Realism: The Best of Both Worlds," *Dialectica* 43:99–124.

van Fraassen, B. (2008) *Scientific Representation*, Oxford UP.

The relationship between intervention and representation is currently resurfacing in philosophy of science. Analytical treatments of the specific intersections between *representation* and *intervention* have recently been explored in Hacking (1983), Radder (2003), Heidelberger (2003), van Fraassen (2008), and Keyser (2017). These accounts analyze intervention-based experimental and measurement practice and the *consequences* for representing and model-building. Of particular interest in my discussion is that some of these accounts explicitly differentiate between representational and productive roles in scientific practice. For example, Heidelberger (2003) and van Fraassen (2008) discuss the representational and productive roles of instruments in experiment and measurement. In the former role, relations in a natural phenomenon are represented in an instrument (van Fraassen 2008, 94). In the latter role, instruments create new phenomena or mimetic phenomena, which resemble natural phenomena. Keyser (2017) takes the distinction between representation and production a step further to differentiate two types of experimental/measurement methodologies:

> When scientists measure/experiment they can *take* measurements, in which case the primary aim is to represent natural phenomena. Scientists can also *make* measurements, in which case the aim is to intervene in order to *produce* experimental objects and processes—characterized as 'effects'. (Keyser 2017, 2)

On Keyser's account 'taking a measurement' involves a scientist using a result in the context of theory to represent a given phenomenon (2017, 9-15). In contrast, 'making a measurement' involves setting up experimental conditions to produce a phenomenon—where that phenomenon can be realized in nature but it can also be a brand new

phenomenon (Keyser 2017, 10). The difference between these two methodologies seems to be a matter of passive representation of a phenomenon vs. active intervention to produce a phenomenon. While the distinction between representation and intervention has been useful in classifying methodology in well-documented contexts like thermometry, microscopy, and cellular measurement, I argue that it falls apart in contexts where taking and making are *entangled*—such as in the context of biomarker measurement in the biomedical sciences.

In this discussion, I aim to show that in *complex methodological contexts*, representational and intervention-based roles require re-conceptualization. I analyze the *relations* between representation and intervention by focusing on the role of intervention in *mediating* representations. In Section 2, I show how applied scientific practice challenges the simple distinction between representational and intervention-based roles of experiment/measurement. In Section 3, I discuss the complex interaction between representation and intervention applied to methodology in biomarker measurement.

2. Methodology at the Intersection between Intervention and Representation

In order to understand why the distinction between representation and intervention needs a multifaceted approach, it is important to be explicit about what it means to represent and intervene in scientific practice. In Section 2.1, I draw on van Fraassen (2008) to discuss representation and both van Fraassen (2008) and Keyser (2017) to discuss intervention. Then in Section 2.2, I show how applied scientific practice challenges the simplistic distinction between representational and intervention-based

roles of experiment/measurement. I argue that the distinction between intervention and representation is less about *specific types of methodologies* in measurement/experiment and more about where one philosophically partitions the measurement *process*.

2.1. Representation and intervention

In experimental and measurement practice, representation has at least three important components: First, instruments or experimental contexts yield measurement values; Second, those values can only be interpreted within the context of a well-developed theory; and third, the relation between the measurement values and the phenomenon is determined by a user (e.g., experimenter). Van Fraassen (2008) provides a rich characterization of representation in measurement and experiment, which requires careful analysis. Worth noting is that van Fraassen takes measurements to be a "special elements of the experimental procedure" (2008, 93-94). For my discussion the embeddedness of measurement in experiment is not important. I will focus on the roles or processes within measurement and experimental practice. But to do this, I will sometimes refer to 'measurement' and other times to 'experiment'. Van Fraassen's characterization focuses on interaction and representation in measurement:

> A measurement is a physical interaction, set up by agents, in a way that allows them to gather information. The outcome of a measurement provides a representation of the entity (object, event, process) measured, selectively, by displaying values of some physical parameters that—according to the theory governing this context—characterize that object. (2008, 179-180)

For van Fraassen, measurement interaction between an object of measurement and apparatus generates a physical outcome—the "measurement outcome" or "physical correlate of the measurement outcome"—, which provides information content about the target of measurement (2008, 143). The contents of measurement outcomes convey information about *what is measured* through the mediation of theory. Van Fraassen posits that theoretical characterization of measurement interaction requires 'coherence':

> The theoretical characterization of the measurement situations is required to be coherent with the claims about the existence of measurement outcomes, their relation to what is measured, and their function as sources of information. (2008, 145)

In short, the theory tells a coherence story about "how its outcomes provide information about what is being measured" (145). Furthermore, the information content is representational. Van Fraassen says, "The outcome provides a representation *of* the measured item, but also represents it *as* thus or so" (2008, 180). To understand how the representational relation works, it is important to refer to van Fraassen's 'representation criterion':

> The criterion for what sorts of interactions can be measurements will be, roughly speaking, that the outcome must represent the target in a certain fashion—, selectively resembling it at a certain level of abstraction, according to the theory—*it is a representation criterion*. (van Fraassen 2008, 141).

Two aspects of the representation criterion require explanation: First, the distinction between "target" and "outcome"; and second, the role of theory in the operation of measurement. I begin with the former. Van Fraassen makes a technical

distinction between the target of measurement ('phenomena') and the outcome of measurement ('appearances'):

> Phenomena are observable, but their appearance, that is to say, *what they look like in given measurement or observation set-ups*, is to be distinguished from them as much as any person's appearance is to be distinguished from that person. (2008, 285)

For van Fraassen, phenomena are observable objects, events, and processes (2008, 283). He emphasizes that phenomena include all observable entities—whether observed or not (2008, 307). A given phenomenon can be measured in many different ways. The outcome of each measurement provides a perspective on a given phenomenon—meaning that the content of measurement tells us what things *look like*, not what they *are like* (2008, 176, 182). The *content* of the measurement outcome is an appearance.

An important qualification is that for van Fraassen, a representation does not represent on its own. The scientist selects the aspects/respects and degrees to which a representation represents a target. This relation can be expressed as: Z uses X to represent Y as F, for purposes P.

Now that the target and outcome of measurement have been characterized, we can specify van Fraassen's role of theory in measurement. According to van Fraassen, "Measurement is an operation that locates an item (already classified as in the domain of a given theory) in a logical space, provided by the theory to represent a range of possible states or characteristics of such items (164). Three things are worth noting about van Fraassen's discussion of logical spaces. First, a logical space provides a multidimensional mathematical space that locates potential objects of measurement (2008, 164). By

measuring we assign the item a location in a logical space. However, according to van Fraassen, it does not have to be on a real number continuum. As van Fraassen points out, items may be classified (by theory) on a range that is "an algebra", "lattice", or a "rudimentary poset" (2008, 172). Second, theoretical location depends on a "family of models" and not just an individual model (2008, 164). Third, an item is located in a "region" of logical space rather than at an exact point (2008, 165). Simply put, theory provides a classificatory system for what is measured. Importantly, theory is *necessary* for this type of classification. Van Fraassen says, "A claim of the form "This is an X-measurement of quantity M pertaining to S" makes sense *only* in a context where the object measured is already classified as a system characterized by quantity M" (2008, 144 my emphasis).

We can summarize the above discussion into four conditions for van Fraassen's account of representation in measurement/experiment practice:

i. Physical Interaction Condition: The interaction between apparatus an object produces a physical correlate of the measurement outcome.

ii. Theoretical Characterization Condition: The content of the measurement outcome is given a location in a logical space, which is governed by a family of theoretical models. An item's location within a logical space can change in content and truth conditions as accepted theories change.

iii. Representational Content Condition: The content of a measurement outcome provides a selective representation of a given target of measurement (phenomenon). Because representations do not represent on their own, users and pragmatic considerations set the representational relation such that: Z uses X to represent Y as F, for purposes P.

iv. Perspectival Information Condition: Measurement generates appearances, which are public, intersubjective, contents of measurement outcomes. Appearances provide selective information about phenomena. Thus information from measurement tells us what something *looks* like and not what something *is* like.

Van Fraassen notes that measurement and experiment are not only limited to a representational role, they can take on at least two productive roles. First, instruments can produce phenomena that "imitate" natural phenomena. That is, carefully controlled conditions give rise to mimetic effects that are used by scientists in the context of theory to resemble natural phenomena (2008, 94-95). It is important to note that van Fraassen emphasizes that natural phenomena are phenomena that exist *independent of human intervention* (2008, 95). The second productive role of instruments is that they are used as "engines of creation" to produce or manufacture new phenomena. Van Fraassen is not explicit about whether or not the representational roles can smear with the productive roles. There is no reason to assume that these roles cannot be combined; but that requires explicit philosophical work to see *how*, which I develop in Section 3.

Keyser (2017) is explicit about the relationship between the representational and intervention-based roles in science. He discusses the *use* of intervention for developing causal representations. Scientists intervene, thereby manipulating causal conditions within a given measurement or experimental system, which he calls 'intervention systems', to produce some sort of "effect" (Keyser 2017, 9-10). According to Keyser, "Intervention systems consist of organized experimental conditions and as such the effects that emerge are often sensitive to changes in conditions*"* (Keyser 2017, 10). Once a given effect is produced it can be used in order to be informative about causal relations for theoretical model building.

Keyser (2017) also differentiates between the methodologies of taking measurements vs. making measurements. I interpret that taking measurements involves three components: First, some instrument or experimental arrangement yields a qualitative or quantitative value; second, a 'theoretical representational framework'—which is just a body of models—is necessary in order to characterize that value according to parameters and relations between parameters; and third, a scientist sets up the resemblance relation between the measurement/experiment value and some aspect(s) of a phenomenon (Keyser 2017, 14-15). In contrast, when scientists make measurements they manipulate causal conditions—such as, preparatory, instrument, and background conditions—within an intervention system. This manipulation gives rise to some effect (Keyser 2017, 3-12).

There is something puzzling about Keyser's distinction between making vs. taking, if we apply the aforementioned conditions (i-iv): i. *Physical Interaction Condition*; ii. *Theoretical Characterization Condition*; iii. *Representational Content*

Condition; and iv. *Perspectival Information Condition*. Namely, it seems that 'making measurements' is compatible with conditions i-iv, so it is not clear why there is a need for a distinction in methodological type, but rather just a difference in details for each condition. For example, when a measurement is made, there is a (i) *physical interaction* that occurs, but it is broader than just the instrument and object. The interaction can include "experimental conditions" (Keyser 2017, 3-5). The product of a made measurement is also amenable to (ii) *theoretical characterization*. Keyser emphasizes that theoretical characterization is necessary for experiment/measurement (Keyser 2017, 14); but he does not make the additional move to say that theoretical characterization is *part of the process* of making a measurement. That is, in order to make a measurement about an effect, one needs to also *characterize* that effect. Without the final characterization, one is only dealing with the material conditions, which is an incomplete part of the measurement process. Keyser can accept that theoretical characterization is a necessary component of making a measurement. Otherwise, he risks offering a limited concept of 'making a measurement' that only applies to arranging the material components of the measurement process and nothing further.

The same challenge goes for (iii) *representational content* and (iv) *perspectival information*. An important component of the measurement process is to represent the relation between the produced effect and some aspect(s) of a phenomenon. For example, is this given effect a limited mimetic representation of a natural phenomenon or is it a brand new phenomenon? Without claims about what the effect is and its relation to objects, events, and processes in the world, 'making a measurement' is uninformative about part of the measurement process: the final value of the measurement outcome.

The aforementioned considerations question the need for a distinction between 'making' vs. 'taking'. One conclusion is that making uses the same components (i-iv), just with slightly different detail. But the other conclusion is a bit unsatisfying: making is really only about organizing the material components, which is an *initial* step in the measurement process, and it does not apply to later steps in measurement.

2.2. Dynamic relations between intervention and representation

I argue that the distinction between intervention vs. representation is less about *specific types of methodologies* in measurement/experiment and more about where to philosophically partition the *measurement process*. To make this point clear, I make two sub-points: 1) Measurement in the biological sciences offers complex and sometimes blurred relations between instrument and object of measurement such that representation and production take on dynamic roles; 2) There is a difference between the act of measurement and the total process of measurement. I briefly describe (1) and (2).

On van Fraassen's (2008) and Keyser's (2017) characterizations of *representation* in measurement, the role of the instrument/apparatus seems to have an important mediating function. It may be the case that philosophical focus on case studies (e.g., thermometry, microscopy, cellular bio, and bacteria) that are instrument-intensive provide a certain support for an instrument-centric account of representation in measurement. Whether or not the necessary mediating role of instruments is an explicit part of both accounts, there is room to develop a richer philosophical view of the role of representation in the total measurement *process*. Without such philosophical development, we risk missing complex cases of measurement where intervention occurs

side-by-side with representation. For example, in some cases of biological measurement, scientists use the organism to measure processes in that same organism but also to represent larger phenomena (Prasolova et al. 2006). For example, mouse diets are manipulated in order to measure chromatin pattern changes. I characterize this as the mouse *constituting experimental conditions* that are being manipulated in order to measure some sort of process. The manipulation of conditions indicates an interventionist approach (or 'making' a measurement). Moreover, without manipulating the mouse's diet scientists would not be able to make a reliable measurement on chromatin structure at all. So the organism is not only being manipulated as part of the experimental/measurement set-up, it is a crucial part of that set-up. That is, without intervention, there is no reliable result. In addition to the organism being used as part of the measurement set-up, it also serves as a physical *representation* of the dynamics of chromatin pattern change. That is, a given model organism can serve as a data model for a specific phenomenon of study—e.g., chromatin pattern in organism X. So, in this case the organism serves a dual function: it constitutes a set of experimental conditions to be manipulated and it serves as a physical representation of a phenomenon. Because of the dual function, this seems to be a case of both 'making' and 'taking' a measurement.

This brings me to sub-point (2). The total process of measurement is often complex in the biological sciences and requires multiple stages of intervening and representing. As mentioned in the model organism example representation and intervention are often *entangled*. Measurement is not merely putting an instrument up to something and waiting for a reading, which can be classified as an *act* of measurement. Measurement is also not merely creating effects out of material conditions. Measurement

requires manipulation of conditions that is *used* in order to generate a representation. For example, identifying a mysterious fungus that is entangled with other fungus in a sample is an active process that requires both intervention and representation. One method is to take a sample and scrape it over a petri dish. What grows are spores that are passively deposited. But if common fungi were commingled with the mysterious fungi in the sample, and the common fungi grew faster, it would be impossible to identify the mysterious fungus. That is, coming back in a couple of weeks and seeing the petri dish covered with familiar species would lead to a false conclusion. Another way to perform the measurement (i.e. culture samples) is as follows. Take the samples and grind them up. Then sprinkle them into a petri dish. Put the dish under the microscope and, using a fine needle, pick out fragments of the mysterious fungus and transplant them to their own dishes (Scott 2010). Once the fragments have been transplanted through this fine-grained intervention, each dish can be left to grow the colonies. The final dishes will offer visual representations that serve as data on the nature of the mysterious fungus. Notice here that intervention is a precursor to reliable representation.

Representation is not only reserved for the final instrument reading. It can also occur at other stages in the measurement process. Likewise, manipulation does not have to occur only at the earlier stages. For instance, organic matter can function as an instrument, like in the case of FourU thermometers, which are RNA molecules that act as thermometers in Salmonella (see Waldminghaus et al. 2007). Suppose that a scientist sets up an experiment to iteratively measure to what extent modifying RNA factors in FourU thermometers changes thermometer readings in Salmonella. In such a case the scientist could modify molecular factors and use the organic thermometers as temperature

measures over many iterations, which would culminate in some sort of data model that organizes the relationship between molecular factors and FourU function. In such a case, there are multiple layers of intervention and representation.

The complex layering of intervention and representation is apparent in biomarker measurement in the biomedical sciences, where biological components serve as representations of disease conditions, but are also intervened on in order to make more reliable representations. I turn to this case study in the subsequent section.

3. Intervening in Representations and Representing Interventions

Biomarkers are used in biomedical measurement to reliably predict causal information about patient outcomes while minimizing the complexity of measurement, resources, and invasiveness. A biomarker is an assayable metric—or simply, an indicator—that is used by scientists to draw conclusions about a biological process (De Gruttola et al. 2001). The greatest utility from biomarker measurement comes from their ability to help clinicians and researchers make conclusions with limited invasiveness. The reliance on biomarkers to make causal conclusions has prompted the use of 'surrogate markers'. These biomarkers are used to substitute for a clinically meaningful endpoint such as a disease condition. A major scientific methodological issue is that the use of multiple biomarkers will produce disagreeing results—and this is true even in the context of biomarkers that use similar biological pathways. To make methodological matters worse, theoretical representation is often not equipped to fill in the causal detail for each biomarker measurement. This amounts to an unfolding methodological puzzle about how to use intervention and representation in biomarkers to produce reliable measurements.

My interest in this case study is not in solving the methodological puzzle, but rather in showing the *relations between intervention and representation* in such a complex case study. In this section, I discuss the complexity of intervention and representation in biomarker measurement to illustrate how intervention mediates the measurement process.

To understand the complex methodology in biomarker measurement it is important to detail the use and limitations of biomarkers. Some biomarkers are used as a substitute for some clinical endpoint. For instance, LDL cholesterol (LDL-C) is a biomarker that clinicians and physicians use to correspond to a clinical endpoint—e.g., heart attack. Moreover, the biomarker is associated with risk factors such as coronary artery stenosis, atherosclerosis, and angina pectoris. Katz (2004) argues that all biomarkers are candidates for 'surrogate markers', which can serve as substitutes for clinical endpoints. That is, surrogate markers are reliable biomarkers that have a one-to-one correspondence with the disease condition such that they can be used to provide reliable predictive and causal information about a given clinical endpoint. There are a couple of points worth noting. First, notice that biomarkers and surrogate markers are being used as representations of a clinical endpoint. That is, to figure out the likelihood of developing a disease condition and to understand the risk factors associated with that disease condition, scientists use biomarkers that indicate information about the endpoint. This means that these physiological components can be used by clinicians and physicians to *represent disease conditions to respects and degrees*. The second point worth noting is that there are many biomarkers but limited surrogate markers and even more limited validated surrogate markers ('surrogate endpoints')—which are surrogate markers that are reliable in multiple contexts of interventions. The importance of this will be relevant

shortly when I discuss the complexity of biomarker measurement. For our purposes, this means that most biomarkers in biomedical practice provide very limited representational information.

Surrogate markers are not passively used as physical representations of disease conditions. Their use is often more effective for representational purposes if there is a *mediating intervention*. For instance, surrogate markers can constitute "response variables". This is where a surrogate marker is manipulated in order to produce an effect that is relevantly similar to the effect with the same manipulation on the clinical endpoint. This means that an adequate surrogate must be "tightly correlated" with the true clinical endpoint; but it also means that any intervention on a surrogate marker must be tightly correlated with the intervention on the true clinical endpoint (Buyse et al. 2000). I interpret this as a dual role for a reliable surrogate marker. It is to act as an epidemiological marker that *represents* some clinical endpoint but also to act as a responding variable that can be used in an *intervention* to causally influence the clinical endpoint. An example of the dual role of the surrogate marker is that high concentrations of LDL cholesterol (LDL-C) correspond to cardiovascular risk (Gofman and Lindgren 1950). But if a therapeutic interventions is used—such as, 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase inhibitors (statins)—that intervention can lower LDL levels, which in turn reduces cardiovascular disease (LaRosa et al. 2005).

So far I have presented the representational and intervention-based role of biomarkers. It is not straightforward to say that surrogate markers are *'made'* like an effect. But it is also not straightforward to say that surrogate markers constitute a *measurement outcome that is the final reading on an instrument*. These markers provide

useful representational information *in the context* of an intervention. To add to the complexity of the relation between representation and intervention, biomarkers in the context of Alzheimer's measurement have added methodological steps. In Alzheimer's measurement there are different biomarkers, which are not correlated with each other and change with independent dynamics in the progression of Alzheimer's disease. So *each* of these biomarkers do not provide the same type of representation about the progression of Alzheimer's disease. Furthermore, scientists *only* understand the disagreement between each of these biomarkers in the presence of different interventions.[1] The different interventions are in the form of drugs (e.g., bapineuzumab and solanezumab) and these interventions produce disagreeing representational results for the biomarkers. That is, the biomarkers respond differently to different interventions, which is methodologically problematic because it indicates that all of these biomarkers cannot be reliably tracking Alzheimer's progression in the same way. Interestingly, scientists systematically compare these disagreeing results to make reliable claims about Alzheimer's progression and treatment (Toyn 2015).[2] To simplify the method used, scientists track how interventions

[1] There has been much work recently on clinical biomarkers like: cerebrospinal fluid (CSF) tau, which is the primary component of neurofibrillary tangles; CSF 42-amino acid amyloid-β (CSF Aβ), which is the protein cleavage product believed to precipitate disease by forming neuron-damaging plaques; and amyloid plaques from PET scans. While the methodological story is beyond the scope of this discussion, there is a complex methodological point that is noteworthy for this discussion (Toyn 2015).

[2] To give a brief picture: The intervention of Bapineuzumab reduces levels of plaque assayed by Aβ PET and CSF tau, but not CSF Aβ; but Solanezumab *does not alter* levels

change properties of biomarkers and then they compare these amalgamated results with how interventions change behavioral/cognitive properties. This type of cross comparison allows scientists to eliminate biomarkers that do not track behavioral/cognitive improvement.

The structure of the methodological complexity in biomarker measurement can be partitioned as follows: 1) For a particular clinical endpoint, there are *limited physical representations* in the form biomarkers (or surrogate markers) which can be *used* to make representational and perspectival conclusions about the endpoint or risk factors associated with it; 2) *Scientists intervene in a process* from each of the biomarkers in order to track the relations between biomarkers and clinical endpoints; and 3) Such interventions prompt *disagreeing results between the biomarkers*, which can 4) be amalgamated by researchers into further representations of the *relations between biomarkers and their clinical endpoints*. The above structural breakdown is merely *a* type of complex methodological process that can occur in biomedical measurement. It shows how interventions on physical representations (biomarkers) can produce other reliable representations. What is important to note about this analysis is the role of intervention in *mediating* further representations. In the case of biomarkers, intervention is necessary to test how close biomarkers are in their representations of clinical endpoints and also to other biomarkers. These representations not only represent the relation between the original biomarker and the clinical endpoint, but they also represent how a given

---

of plaque assayed by Aβ PET and CSF tau but leads to a *reduction in* CSF Aβ. Cross comparison of the *intervention* mechanisms allows scientists to begin to make causal claims about which biomarkers are more reliable than others (Toyn 2015).

intervention affects a given biomarker. As such, intervention paves the way for iterations of representations.

4. Concluding Remarks

In this discussion, I have analyzed the role of intervention in mediating representations by using examples from the biological and biomedical sciences. Characterizing intervention as a mediating factor in a larger methodological operation provides an important point about scientific practice. Representation and intervention are not neatly partitioned into contrasting methodologies. In fact, applied science often dictates the complex, and often smeared, philosophical concepts and methodologies. For this reason, I am proposing a *process* view of intervention and representation. This view opens up the diversity of relations between representation and intervention in a given experimental/measurement practice. While I have emphasized how intervention mediates representation, there is more territory to explore about the mediating role of representation for intervention.

Work Cited

De Gruttola, V.G, Clax P, DeMets DL, et al. (2001). Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. Control Clin Trials 22:485–502.

Gofman, J.W., Jones, H.B., Lindgren, F.T., et al (1950). Blood lipids and human atherosclerosis. *Circulation* 2:161–178.

Hacking, I., (1983). *Representing and Intervening*, Cambridge: Cambridge University

Press.

Heidelberger, M. (2003). Theory-ladenness and scientific instruments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 138–151). Pittsburgh, PA: University of Pittsburgh Press.

Katz, R. (2004). Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* 1:189–195. doi: 10.1602/neurorx.1.2.189

Keyser, V. (2017). Experimental Effects and Causal Representations. *Synthese*, SI: Modeling and Representation, pp. 1-32.

LaRosa, J.C., Grundy, S.M., Waters, D.D., et al. (2005). Intensive Lipid Lowering with Atorvastatin in Patients with Stable Coronary Disease. *New England Journal of Medicine* 352:1425–1435. doi: 10.1056/NEJMoa050461

Prasolova L.A., L.N. Trut, I.N. Os'kina, R.G. Gulevich, I.Z. Pliusnina, E.B. Vsevolodov, I.F. Latypov. (2006). The effect of methyl-containing supplements during pregnancy on the phenotypic modification of offspring hair color in rats. *Genetika*, 42*(1)*, 78-83.

Radder, H. (2003). Technology and theory in experimental science. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 174–197). Pittsburgh, PA: University of Pittsburgh Press.

Toyn, J. (2015). What lessons can be learned from failed Alzheimer's disease trials? *Expert Rev Clin Pharmacol* 8:267–269. doi: 10.1586/17512433.2015.1034690

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.

Waldminghaus, T., Nadja H., Sabine B., and Franz N. (2007). FourU: A Novel Type of

RNA Thermometer in Salmonella. *Molecular Microbiology* 65 (2): 413–24. https://doi.org/10.1111/j.1365-2958.2007.05794.x.

Philosophy of Science (forthcoming)
v1.2 (as of 9/15/18)
Please cite published version

Are Emotions Psychological Constructions?

Charlie Kurth
Department of Philosophy
Western Michigan University

Abstract: According to psychological constructivism, emotions result from projecting folk emotion concepts onto felt affective episodes (e.g., Barrett 2017, LeDoux 2015, Russell 2004). Moreover, while constructivists acknowledge there's a biological dimension to emotion, they deny that emotions are (or involve) affect programs. So they also deny that emotions are natural kinds. However, the essential role constructivism gives to felt experience and folk concepts leads to an account that's extensionally inadequate and functionally inaccurate. Moreover, biologically-oriented proposals that reject these commitments are not similarly encumbered. Recognizing this has two implications: biological mechanisms are more central to emotion than constructivism allows, and the conclusion that emotions aren't natural kinds is premature.

This paper challenges the psychological constructivist account of emotions that is gaining prominence among neuroscientists and psychologists (e.g., Barrett 2017, 2012, 2009; LeDoux 2015; Russell 2004). According to constructivism, emotions result from projecting culturally-fashioned concepts onto felt affective episodes. Fear, for instance, just is a feeling of negative arousal as viewed through the lens of one's folk concept FEAR. This proposal is novel in taking felt experience and cognitive projection to be essential elements of what emotions are. Moreover, while constructivists acknowledge that there's a biological dimension to emotions (e.g., neural mechanisms are responsible for generating the conscious feelings that we project our emotion concepts on to), they deny that emotions are, or necessarily involve, anything like an affect program. Thus, constructivism is philosophically significant in two ways. First, in denying an essential role for biological mechanisms, it challenges influential, affect-program-oriented accounts of emotion (e.g., Scarantino & Griffiths 2011; Ekman & Cordaro 2011). Second, in understanding emotions as projections of folk emotion concepts, it takes emotions to be social-psychological constructions, not natural kinds.

But despite constructivism's appeal among cognitive scientists, the role that it gives to felt experience and folk concepts leads to an account of emotion that's both extensionally inadequate and functionally inaccurate. Moreover, biologically-oriented proposals that reject constructivism's problematic commitments are not similarly encumbered. Recognizing all this reveals that an adequate account needs to give greater place to the biological mechanisms that underlie emotions than constructivism allows. This, in turn, suggests that the constructivists' conclusion that emotions are not natural kinds is premature.

1. Psychological Constructivism and Its Appeal

Constructivism sees emotions as having two elements: a felt affective experience and a cognitive projection or labeling. Taking these in turn, the felt experience component—or "core affect" as it's often called—is a neurophysiological state that manifests as a consciously experienced combination of valence (i.e., feeling good or bad) and arousal (i.e., feeling activated or deactivated) (Barret 2006: 48; Russell 2004; LeDoux 2015: 226-232). Importantly, constructivism's focus on core affect looks just to the amalgamated *experience* of these two components—valence and arousal. What *causes* this felt experience is irrelevant to the nature and individuation of emotions. In fact, and as we will see, allowing that particular sensations (instances of core affect) can be produced by a range of distinct neural circuits or somatic events is taken to be a point in favor of the constructivist proposal.

Given this account of the felt dimension, constructivism maintains that "discrete emotions emerge from a conceptual analysis of core affect. Specifically, the experience of feeling an emotion…occurs when conceptual knowledge about emotion is brought to bear to categorize a momentary state of core affect. … [These] [c]ategorization processes enact the rules, [that guide] the emergence of an emotional episode" (Barrett 2006: 49; also LeDoux 2015: 225-232). This talk of "conceptual analysis," "conceptual knowledge," and "categorization" should be understood thinly.

The underlying process needn't involve some full-fledged, conscious judgment. Rather, all that's necessary is an unconscious or implicit recognition that one's sense of one's situation, and one's felt physiological state, fall under a particular folk emotion concept.

These emotions concepts, in turn, should be understood as folk theories or culturally-shaped behavioral scripts that detail the nature and function of the particular mental states picked out by specific emotion labels ('fear,' 'joy,' 'anger,' etc.). Moreover, the fact that folk emotion concepts engage these folk theories and behavioral scripts entails that the projecting of a particular label onto an instance of core affect not only imbues one's situation with the associated, emotionally-colored meaning, but also shapes one's subsequent thoughts, physiological responses, and behaviors (Barrett 2012; LeDoux 2015).

Formalizing this a bit, we can see psychological constructivism as committed to four theses:

> (PC1) Each emotion type/category is constituted by the projecting of a specific folk emotion concept (e.g., FEAR, JOY) onto a felt affective experience.
>
> (PC2) Token emotion episodes (e.g., a given instance of fear) are cognitive acts where one (implicitly) labels an occurrent conscious feeling with a particular folk emotion concept and so comes to see the feeling through the lens of that concept.
>
> (PC3) There is no unique (set of) neural circuit(s) or psychological mechanism(s) responsible for the conscious feelings that get categorized with particular folk emotion concepts.
>
> (PC4) The act of labeling a feeling with a particular folk emotion concept affects one's subsequent thoughts, physiological responses, and behaviors.

According to its advocates, much of constructivism's appeal lies in its explanatory power. In comparison to more biologically-oriented theories, it provides a better explanation of empirical research on the biological mechanisms and correlates associated with emotions (e.g., neural circuits, patterns of physiological change, and expressive behavior). Since the discussion that follows will build

from the contrast between constructivism and competing biologically-oriented theories (BTs), it will be useful to briefly sketch the BT approach and the constructivists' case against it.

As a generalization, BTs maintain that emotions are, or necessarily engage, affect programs—that is, largely encapsulated systems that automatically prompt stereotyped patterns of physiological changes, expressive behavior, motor routines, attentional shifts, and forms of higher-cognitive processing in response to (evolutionarily-relevant) threats and opportunities. So, for example, fear is (or essentially involves) an affect state that consists of automatically engaged tendencies for inter alia increases in arousal, narrowing of attention, and the cueing of fight/flight/freeze behavior in response to the perception of some danger.

But since BTs take affect programs to be essential (even identical) to emotions, constructivists argue they cannot explain two well-documented sets of findings.[1]

> (F1) One can feel a given emotion without engaging what science suggests is the best candidate for its underlying biological drivers (or their correlates)—e.g., activation of particular neural circuits, a distinctive physiological response, characteristic expressive behavior.
>
> (F2) The relevant biological drivers/correlates can be engaged though one does not report feeling the associated emotion.

So, for instance, though the central nucleus of the amygdala (CeA) is thought to be central to fear, research shows both that individuals will report being afraid when the CeA is not engaged (F1), and that the CeA can be active though individuals report not feeling fear (F2).

BT proponents have sought to address these explanatory limitations by insisting that we must narrow our understanding of what, say, FEAR is. More specifically, they maintain that the folk emotion concepts that the above research relies on (in, e.g., the self-reports of emotions (not) felt) are too

[1] See, e.g., Barrett 2012 for a review of the relevant empirical work.

coarsely grained for scientific investigations like these. The BT advocates' expectation is that a more refined account of what 'fear' refers to will reduce, even eliminate, dissociations of the sort noted above (e.g., Scarantino & Griffiths 2011; Kurth 2018). But constructivists respond that any effort to narrow or otherwise refine our emotion concepts along these lines will result in an account of (e.g.) fear that is troublingly stipulative or excessively revisionary with regard to our ordinary understanding of these emotions (Barrett 2012: 415-6; LeDoux 2015: 234).

Two aspects of this debates are particularly important for our purposes. First, central to the constructivist complaint is the move to take a failure to accommodate our *ordinary emotion talk* as the standard for what counts as stipulative or excessively revisionary account. Second, given our ordinary emotion talk as the standard, the above four theses appear to give constructivism the resources and flexibility it needs to explain not just (F1)-(F2), but also the richness and cultural variation of emotional life more generally (e.g., Barrett 2012, 2009). However, I will argue that investigating the extensional adequacy and functional accuracy of constructivism's core theses provides us with reason to doubt each of (PC1)-(PC4).

2. Is Constructivism Extensionally Adequate?

As we've seen, a central feature of the debate between constructivism and BTs is the charge that BTs cannot accommodate dissociation data without committing to a stipulative or excessively revisionary account of what emotions are. In what follows, I give three examples that suggest constructivism faces a similar problem. More specifically, a closer look at the constructivists' dual claim that emotions are *cognitive labelings* of *felt experiences* reveals that the account is both under- and over-inclusive with regard

to our ordinary understanding of things like: what emotions are, when we experience them, and how they differ from moods, feelings, and other categories of affect.[2]

First consider the constructivist's commitment to understanding emotions as felt experiences—that is, changes in core affect that we're consciously aware of. An implication of taking felt affective experience as essential to being an emotion is that it rules out the possibility of unconscious emotions. Some constructivists appear to embrace this result. For instance, Joseph LeDoux maintains that claims about unconscious emotions are "oxymoronic" (2015: 234; also, 19). But LeDoux's acceptance of this implication aside, the thought that there cannot be unconscious emotions fits poorly with our everyday experiences and our ordinary emotion talk.

For instance, if there aren't unconscious emotions, then how do we explain situations where we don't realize that we were (say) afraid until *after* the danger has passed? Pressing further, notice that we not only regularly speak of unconscious emotions, but also appeal to them in order to explain our behavior. For example, we say things like, "Bill won't discuss the book he is working on. He says it's not ready yet—but he doesn't realize that he's really just afraid about getting negative feedback." While ordinary talk like this is easy to make sense of on the assumption that Bill is unconsciously fearful, such an explanation isn't available to a constructivist like LeDoux—our ordinary talk to the contrary, Bill isn't unconsciously afraid, but rather experiencing some other psychological blockage.

But the constructivists' trouble with unconscious emotions runs deeper—the case for their existence also has empirical support. For instance, recent experimental work has shown that subliminally presented emotion faces can produce affective responses that bring emotion-specific behaviors *even though* the subject denies feeling an emotion. In particular, subliminally presented happy

[2] Thus the strategy I employ here—one that *grants* constructivists' their criterion for assessing when an account is excessively revisionary—is distinct from standard defenses of BTs noted in §1.

faces bring increased "liking" behavior (e.g., greater consumption of a novel beverage), while subliminally presented angry faces have the opposite result (Winkielman et al. 2003; also, Kihlstrom 1999). Since these patterns of behavior mesh with our understanding of both joy as an emotion that tends to increase interest/engagement, and anger as an emotion that brings avoidance/rejection tendencies, these results are taken as evidence of unconscious emotions.

While the constructivist might try to pass these findings off as cases where unconscious changes in core affect (not emotion) produce the behaviors, the plausibility of the proposal is undercut by the fit we find between the subliminally presented happy (angry) face, the resulting liking (avoidance) behavior, and *our ordinary understanding* what happiness (anger) involves (Winkielman et al. 2005). The upshot, then, is that constructivism's insistence that felt changes in core affect are *essential* to what emotions are has revisionary implications with regard to our ordinary (and scientific) understanding of emotional life.

But even if we're willing to grant that our talk of unconscious emotions is merely metaphorical—an elliptical way of talking about some non-emotion form of (unconscious) affect—the constructivist's second core commitment brings additional problems. In particular, the claim that emotions are the product of our cognitive labelings/projections makes facts about when we are experiencing an emotion—and what emotion it is—too sensitive to random situational features and framing effects. To draw this out, consider the following case.

> Coffee. I order a cup of decaf coffee and sit down to read a magazine cover story about Trump's latest foreign policy provocations. But unbeknownst to me, the barista confuses my order and I get a cup of regular coffee. As the caffeine works its way into my system, it brings a (consciously experienced) change in my arousal. As a result, I start reading the article with jittery attentiveness.

Given the scenario, it seems my jittery, attentive reading is best understood as a bout of caffeine-induced hyperactivity. But notice: there's nothing in the constructivist account to rule out the

possibility that I'm actually having an emotional experience—I'm afraid. After all, on the constructivist account, this experience could be a change in core affect that I've (implicitly) labeled 'fear.' While that possibility alone seems odd (to my ear, at least, the case is best understood as emotionless hyperactivity, not fear), there's more trouble.

To draw this out, consider the constructivist's likely response to the case. Given the setup, she would likely maintain that whether this is an instance of fear depends on whether I see it that way—what sort of meaning do I attribute to my situation (e.g., Barrett 2017: 126; 2012: 419-420; 2009: 1293)? For instance, if I assent to the barista's remark that I seem really uneasy about the article that I'm reading, then—by (implicitly) labeling my behavior through my assent—I imbue my situation with the meaning carried by my FEAR concept. I am, therefore, feeling fear. While this move might seem to allow the constructivist a way to account for the case, it comes at a high cost. For notice, had the barista instead said something like, "Whoops, I messed up and gave you regular, not decaf—no wonder you're so hyper," I'd likely assent to that too. And so I wouldn't be afraid—just hyperactively aroused.

But that's odd. Our ordinary thinking about emotions suggests that whether I'm experiencing a particular emotion, and what emotion I'm experiencing, should *not* be so sensitive to random situational features like what questions the barista—or anyone for that matter—just happen to ask me. To be clear, the claim here is not that emotions are immune to situational and contextual factors. Rather, the point is that on the constructivists' account emotions turn out to be *too* sensitive to them. The radical situational sensitivity entailed by constructivism makes it not only too easy to experience an emotion, but also ties facts about what emotion we're experiencing to irrelevant situational factors.

Together, the difficulties raised by unconscious emotions and incidental situational features call the extensional adequacy of the constructivist account into question and do so in a way that

pinpoints the commitments of (PC1) and (PC2) as the source of the trouble—after all, these claims posit feelings of core affect and projections of folk concepts as essential to what emotions are. Of equal note is the fact that biological theories are less vulnerable to these difficulties. For one, irrelevant situational features should have less influence on what emotion one happens to experience since, according to BTs, emotions are (or are principally driven by) affect programs, not contextualized cognitive labelings. Moreover, since affect programs are things that can operate below that level of conscious awareness (Kurth 2018), taking emotions to be driven by affect programs provides BTs with the resources needed to explain unconscious emotions.

While the above discussion raises worries about the first two constructivist theses (PC1-PC2), it also provides the makings for worries about the third. In particular, because constructivism denies (via PC3) that emotions are underwritten by affect programs, it has trouble making plausible distinctions between emotions and similar states like moods. To draw this out, notice that the coffee case from above can be easily extended to show that constructivism makes it too easy to flip between moods and emotions. All we need to do is substitute "being in a worried mood" for "hyperactive" in the presentation of the case. Once we do this, we see that mere changes in the question the barista asks me can change whether I'm worried (a mood) or afraid (an emotion).

So we again see that constructivism has problematic explanatory limitations—this time with regard to preserving the thought that there's a substantive difference between moods and emotions. On the constructivist account, this distinction is just a matter of how we happen to label our felt experiences. While some constructivists appear willing to accept this conclusion (e.g., Barrett 2017, 2009), it highlights another place where the constructivist proposal has revisionary implications—after all, moods and emotions are generally thought to be *distinct* forms of affect (e.g., Ben-Ze'ev 2000: Chap. 4). Moreover, here too we have a difficulty that's easily avoided by biological accounts. Since

BTs take emotions to be (driven by) affect programs, they can appeal to the engagement of these mechanisms as the basis for the emotion/mood distinction (e.g., Kurth 2018; Wong 2017).

Stepping back, then, although constructivism purports to be less stipulative with regard to capturing our ordinary understanding of emotions, the above examples call this into question. For starters, the constructivists' commitment to (PC1)-(PC3) has revisionary implications for our ordinary understanding of what emotions are, when we experience them, and how they differ from moods. Moreover, we have also seen that biologically-oriented accounts—in eschewing this trio of problematic theses—are better equipped to provide a plausible account of these features of our everyday emotion talk.

3. Is Constructivism Functionally Accurate?

The challenges to the constructivist picture extend beyond concerns about its extensional adequacy. The account also makes predictions about how projecting emotion concepts onto felt experience should shape subsequent behavior that are poorly supported by the empirical record. Two examples will draw this out.

First consider emotion misattribution research. In this work, a feeling that is typically associated with a particular emotion (e.g., feelings of unease and anxiety) is subtly induced, but the individual is lead to believe they are not, in fact, experiencing that emotion but rather something else (e.g., the effects of caffeine). Constructivism predicts (via PC4) that individuals in these experiments should display different behaviors depending on whether they are in the control or misattribution conditions. For instance, individuals led to believe that the unease they're feeling is not anxiety, but something else (caffeine) should display diminished anxiety-related behaviors in comparison to controls who were not misled about their unease. But on this score, the experimental findings are decidedly mixed.

First, while there is a sizable body of findings showing misattribution manipulations attenuate subsequent emotion-related behavior, there is also a sufficiently large set of non-confirmations to raise concerns. For instance, while some research on public speaking anxiety suggests that attributing unease to a pill you just took rather than anxiety about a public talk you must give leads to a reduction in anxiety-related behaviors—stuttering, apprehension, and the like (Olson 1988), other studies have failed to find any differences in these behaviors (Slivkin & Buss 1984; Singerman, Borkovec & Baron 1976).

Moreover, even in cases where emotion-related behavior is reduced in the manipulation condition, it's not clear how much support this brings to the constructivist. This is because it's often unclear whether the reductions in emotion-specific behavior are (i) the result of the misattribution or (ii) a consequence of directing subjects' attention away from the emotion eliciting stimuli (for a review, see, e.g., Reisenzein 1983). This potential confound is problematic for constructivists since only possibility (i) provides direct support for the claim of (PC4)—namely, that the act of labeling *itself* affects subsequent behavior.

The second problematic set of results comes from work in political science. This research investigates how negative emotions shape public policy decision making among voters (e.g., MacKuen et al. 2010; Brader et al. 2008; Valentino et al. 2008). The core hypothesis of this research is that negative emotions (especially, anger and anxiety) affect subsequent behavior in different ways. In particular, anger—as a response to challenges to what one values—should tend to bring behavior geared toward defending the threatened values. By contrast, since anxiety is a response to uncertainty, it should tend to bring caution and information gathering aimed helping one work through the uncertainty one faces.

To test these predictions, the experimental set up works as follows. First, individuals are asked to read a (fake) news story designed to provoke anger or anxiety by challenging the individuals' pre-existing views about contentious policy issues like immigration, affirmative action, and economic policy. After reading the story, the participants are given the opportunity to use a website containing links to additional information, both for and against, the policy issue at hand. They are also asked how the original news story they read made them feel (e.g., angry, anxious). So by tracking what kinds of information the participants looked at through the website, experimenters can identify differences in how the anger and anxiety provoked by the story shaped subsequent behavior.

In the present context, these experiments allow us to test a pair of predictions that follow from the constructivist theses (PC1) and (PC4):

> (P1) Labeling felt experiences with distinct folk emotion concepts should bring different patterns of behavior.
>
> (P2) The behaviors that result from labeling a felt experience with a particular concept should map to our folk understanding of the emotion in question.[3]

More specifically, given (P1) and (P2), we should see different behaviors based on whether the participants in the experiment label their emotion 'anger' or 'anxiety' (P1). Moreover, the different behaviors should map to the above, ordinary understanding of these emotions—e.g., angry individuals should look for information that helps them defend their preferred policy position, while anxious individuals should engage less in motivated inquiry and more in open-minded forms of investigation (P2).

[3] As evidence of constructivism's commitment to these predictions, consider Lisa Feldman Barrett's comment that "when a person is feeling angry…she has categorized sensations from the body and the world using conceptual knowledge of the category 'anger'. As a result, that person will experience an unpleasant, high arousal state as evidence that someone is offensive. In fear…she will experience the same state as evidence that the world is threatening. And, *either way, the person will behave accordingly*" (2009: 1293, emphasis added).

However, whether we find support for these predictions turns—surprisingly—on what the policy issue used in the experiment was. More specifically, in experiments where the policy question that was challenged by the fake news story concerned immigration, the results fit poorly with constructivism's predictions. That is, participants behaved in the same angry way regardless of whether they reported feeling anger or anxiety (Brader et al. 2008). By contrast, if the policy issue at hand concerned affirmative action or economic policy, the results are more in line with (P1)-(P2): anger and anxiety provoked by the news stories not only brought different patterns of behavior, but the resulting behaviors mesh with our ordinary conception of how these emotions function (MacKuen et al. 2010; Valentino 2008).

While this second set of results might appear to be good news for constructivists, the trouble lies in explaining why we get the different results between the immigration and affirmative action/economic policy experiments. After all, other than the content of the issue at hand, the experimental designs were *identical*. In response, the constructivist might argue that content and context matter (e.g., Barrett 2012, 2009): the similar behaviors that subjects display in the immigration version of the study suggest that the cultural scripts associated with 'anger' and 'anxiety' are highly sensitive to negative stereotypes about minorities. More specifically, the thought would be that there's something about the combination of immigration debates and racial stereotypes that changes the standard behavioral scripts associated with 'anger' and 'anxiety' so that, while they *typically* generate different behaviors, they *now* bring the same ones.

But setting aside concerns about the ad hoc nature of this proposal, without more of a backstory, it's unconvincing. After all, affirmative action debates are *also* framed in racial stereotype provoking ways. So here too we should see anger and anxiety generating similar patterns of behavior. But we don't.

Moreover, notice that, on this front, biological accounts have an easier time explaining the experimental findings. For instance, as one possibility, the BT advocate could argue that only participants in the immigration study are likely to be experiencing *both* anger and anxiety: anger about the harms immigrants will bring and anxiety given their uncertainty about the likelihood of these harms. Given this, the BT advocate could then add two claims about what happens when both these emotions are engaged. First, since anger is a more powerful emotion than anxiety, it tends to win out with regard to shaping individuals' subsequent behavior. Second, given the high degree of overlap in the felt experiences produced by the anger and anxiety affect programs (e.g., both bring increased, negatively valenced arousal), when prompted to state what emotion they are feeling, some subjects happen to interpret their feelings as anger, while others see it as anxiety. Thus, the BT advocate can explain both why we get mixed results when subjects are prompted to state what emotion they are feeling and why, despite these differences in self-reports, the individuals nonetheless respond with behavior characteristic of anger, not anxiety. Moreover, because this proposal allows anger to drive behavior *regardless* of how subjects happen to label it, the explanation is unavailable to constructivists.

All told, we have two independent sets of experimental findings showing (at best) equivocal support for constructivism's predictions about how projecting emotion concepts onto felt experience should shape subsequent behavior. Moreover, we've also learned that more biologically-oriented accounts are better able to handle the experimental findings we've reviewed.

4. Conclusion: Emotions, Biology, and Natural Kinds

As we've seen, constructivism's purported advantage over more biologically-oriented theories lies is its ability to better explain the richness and diversity of emotional life (§1). But we have also seen that a crucial premise in this argument is the move to take accommodating our ordinary emotion talk as the standard for assessing a theory's explanatory power. Not only are there familiar problems for

adopting such a standard (e.g., Scarantino & Griffiths 2011, Kurth 2018), but—even if we accept it—we've learned that there's trouble for constructivism. In particular, the explanatory "success" constructivism secures come by way of a highly revisionary account of what emotions are, when we experience them, how they differ from moods, and the way that they shape behavior (§§2-3). Moreover, our critical observations also implicate the four constructivist theses (PC1-PC4) as the source of these difficulties. Thus it's not surprising that more biologically oriented proposals—accounts that reject these commitments—do not face similar explanatory limitations.

Taken together, then, the arguments of this paper suggest a pair of larger lessons. First, even if we agree that constructivists are correct about what the relevant standard for assessing a theory of emotion is, we've learned that an adequate account must give greater place to the biological mechanisms that underlie emotions than constructivism allows. This, in turn, indicates that the constructivists' conclusion that emotions are not natural kinds is premature. After all, if we must posit something like an affect program in order to (i) explain everyday talk and empirical findings about unconscious emotions, (ii) capture the thought that emotional experience is not radically sensitive to random situational features, and (iii) accommodate research regarding how emotions shape behavior, then we have evidence that (at least some) emotions are underwritten by mechanisms that make them plausible candidates for being natural kinds.

References

Barrett, L. 2017. *How Emotions Are Made*. New York: Houghton Mifflin Harcourt.

---. 2012. "Emotions Are Real." *Emotion* 12: 413-429.

---. 2009. "Variety is the Spice of Life." *Emotion and Cognition* 23: 1284-1306.

---. 2006. "Emotions as Natural Kinds?" *Perspectives on Psychological Science* 1: 28-58.

Ben-Ze'ev, A. 2000. *The Subtlety of Emotions*. Cambridge.

Brader, T. et al. 2008. "What Triggers Public Opposition to Immigration?" *American Journal of Political Science* 52: 959-978.

Ekman, P. & D. Cordaro. 2011. "What is Meant by Calling Emotions Basic." *Emotion Review* 3: 364–370

Kihlstrom, J.F. 1999. "The Psychological Unconscious." In L.A. Pervin & O.P. John (Eds.), *Handbook of Personality* (2nd ed., pp.424–442). New York: Guilford Press.

Kurth, C. 2018. *The Anxious Mind*. MIT Press.

LeDoux, J. 2015. *Anxious*. New York: Viking.

MacKuen, M. et al. 2010. "Civil Engagements," *American Journal of Political Science* 54: 440–458.

Olson, J. 1988. "Missattribution, Preparatory Information, and Speech Anxiety" *Journal of Personality and Social Psychology* 54: 758-767.

Reisenzein, R. 1983. "The Schachter Theory of Emotion" *Psychological Bulletin* 94: 239-264.

Russell, P. 2004. "Core Affect and the Psychological Construction of Emotion. *Psychological Review* 110: 145–172

Scarantino, A & P. Griffiths. 2011. "Don't Give Up on Basic Emotions" Emotion Review 3: 1-11.

Singerman, K. et al. 1976. "Failure of a 'Misattribution Therapy' Manipulation with a Clinically Relevant Target Behavior" *Behavior Therapy* 7: 306-316.

Slivken, K. & A. H. Buss. 1984. "Misattribution and Speech Anxiety" *Journal of Personality and Social Psychology* 47: 396-402.

Valentino, N. et al. 2008. "Is a Worried Citizen a Good Citizen?" *Political Psychology* 29: 247–73.

Winkielman, P. et al. 2005. "Unconscious Affective Reactions to Masked Happy versus Angry Faces Influence Consumption Behavior and Judgments of Value." *Personality and Social Psychology Bulletin* 121-135.

Wong, M. 2017. "The Mood-Emotion Loop" *Philosophical Studies* 173: 3061-3080.

Symposium: *Bridging the Gap Between Scientists and the Public*, PSA 2018

How trustworthy and authoritative is scientific input into public policy deliberations?[i]

Hugh Lacey
Swarthmore College / University of São Paulo

Abstract: Appraising public policies about using technoscientific innovations requires attending to the values reflected in the interests expected to be served by them. It also requires addressing questions about the efficacy of using the innovations, and about whether or not using them may occasion harmful effects (risks); moreover, judgments about these matters should be soundly backed by empirical evidence. Clearly, then, scientists have an important role to play in formulating and appraising these public policies.

However, ethical and social values affect decisions made about the criteria (1) for identifying the range of risks, and of relevant empirical data needed for making judgments about them, that should be considered in public policy deliberations, and (2) for determining how well claims concerning risks should be supported by the available data in order to warrant that they have a decisive role in the deliberations. Consider the case of public policies about using GMOs. Concerning the range of data: is it sufficient for risk assessment only to be informed by data relevant to investigating the risks of using GMOs that may be occasioned by way of physical/chemical/biological mechanisms directly triggered by events within their modified genomes? Or: should data pertaining to the full range of ecological and socioeconomic effects of using them, in the environments in which they are used and under the socioeconomic conditions of their use, also inform this assessment? Those interested in producing and using GMOs, in the light of their adhering to values of capital and the market, are likely to give a positive answer to the first question; those holding competing values, e.g., connected with respect for human rights and environmental sustainability, to the second. And, concerning the degree of support: the former – citing the ethical gravity of losses (both economic and, allegedly, for food security) that would be incurred by failing to use GMOs on a wide scale – are likely to require less stringent standards of evidential appraisal than the latter.

Scientists, *qua* scientists, however, do not have special authority in the realm of values. Thus, their judgments, about the evidential support that claims about risks (and some other matters) have, may sometimes be reasonably (although not decisively) contested partly on value-laden grounds – as they have been in the GMO case, where the contestation has generated considerable controversy, and continues to do so. It follows that, in the context of deliberations about public policy, unless scientists engage with representatives of all stakeholders in the outcomes of the policies (as, for the most part, has not happened in the GMO case) – taking into account that their competing values may lead to making different decisions about what are the relevant data, as well as about the degree of support required for their claims about risks to gain the required credibility to inform the deliberations; and respecting "tempered equality" of participants in the dialogue (Longino) – their trustworthiness is put into question and their authority diminished.

1.

In a letter, dated June 29th, 2016, 135 Nobel laureates made the following claims, among others,[ii] related to using GMOs (genetically modified organisms) in agriculture:

(i) "Scientific and regulatory agencies around the world have repeatedly and consistently found crops and foods improved through biotechnology to be as safe as, if not safer than those derived from any other method of production."

(ii) "There has never been a single confirmed case of a negative health outcome for humans or animals from their consumption."

(iii) "Their environmental impacts have been shown repeatedly to be less damaging to the environment, and a boon to global biodiversity" (Laureates Letter, 2016).

Reflecting the authority and esteem that tends to be accorded to Nobel laureates, the declaration was widely reported and taken to bolster the allegation that there is a *scientific consensus* that cultivating and harvesting genetically engineered crops, and consuming their products, is safe.[iii] The scientists who signed it aimed to assure the public that the three claims are well con-

firmed, and that public policy and regulatory deliberations should reflect them. The claims do not derive from outcomes of the research conducted by these scientists, for at most one or two of them (so far as I can tell, none) have themselves engaged in biosafety research. They were putting their authority behind the research and judgments of others, whom presumably they trusted. Even so, one might reasonable assume that they had, before signing the declaration, examined the relevant research and concurred with its outcomes, and had found good reason to tell us, as they do, (presumably based on a thorough examination of its writings and actions) that the opposition is "based on emotion and dogma contradicted by data" and that it "must be stopped." At the end of the paper, I will argue that the declaration misuses scientific authority and contributes to doubts about the trustworthiness of leading scientific authorities. My larger purpose, however, is to suggest **some** necessary conditions for re-establishing trust in scientific communities – bridging the gap between scientists and the public, and (the concern of de Martín-Melo & Intemann, 2018) – so that both the authority and integrity of science, and the conditions for strengthening democratic societies, are enhanced

2.

First, some more general remarks. I maintain that the deliberations out of which arise public polices having to do with introducing, using and regulating technoscientific innovations (I only have time to discuss GEOs) should consider:

(1) questions about the *efficacy* of the proposed uses are addressed – and about their *safety*, specifically about how well available empirical evidence confirms that the proposed uses do not occasion harmful effects (or risks of causing harmful effects);

(2) the values reflected in the interests expected to be served by the proposed uses, as well as questions about whether interests expected to be served by competing values may be disadvantaged by them, and priorities among the competing interests;

(3) identified potential alternatives to using these innovations – including fundamentally different kinds of practices – as well as how using them compares to the proposed uses with respect to efficacy and safety (and other potential benefits).[iv]

Of these conditions only (1) is uncontroversial and generally followed (although there are disagreements about how it ought to be followed) in public policy deliberations.[v] Clearly satisfactory answers to the questions about efficacy and safety depend on trustworthy and reliable scientific input. I will not question that scientific research has reliably established the efficacy of the GEOs that have already been approved by regulatory bodies for agricultural use, for the most part GEOs with herbicide-resistant and insecticidal properties.[vi] Efficacy does not imply safety, however, and the research approaches (in molecular biology, biotechnology, etc) within which efficacy is established do not suffice for engaging in research dealing with safety. However, many regulatory practices presuppose that scientific input, pertaining to deliberations about safety – like that about efficacy – is obtained prior to consideration of (2) and (3), and to entanglement with value questions. Hence, the currency of the terms "scientific risk assessments" and

"scientific safety studies", areas of research in which scientific/technical "experts" should be granted authority.

One needs to be wary here, for "safe" and "risk" are 'thick ethical terms'. Scientific safety studies cannot be fully separate from entanglement with values and obligations. Thus, e.g. (simplifying a little), 'using X *is* unsafe' implies (*ceteris paribus*) 'X *should* not be used, unless appropriate precautions are taken.' And, when scientists conclude, on the basis of their investigations, that 'using X is safe', they intend it to follow (and to have impact at step (2)), that *ceteris paribus* 'it is improper to impede using X'.[vii] This does not mean that, in the course of empirical research in scientific safety studies, value-laden terms are used in articulating hypotheses and reporting empirical data. The link between the results of the empirical research and the subsequent value judgments depends on a step (call it step (0)), casually made prior to the empirical investigations. At step (0), the set of possible unintended collateral effects of using X is scrutinized, and those that are identified as harmful (as risks)[viii] – obviously value judgments are made here – are then investigated for such matters as the probability and magnitude of their possible occurrence, and its being countered by introducing scientifically informed regulations. In the investigation, the possible collateral effects are characterized, not with thick ethical terms, but with theoretical and observational terms deployed in relevant scientific fields, like molecular biology, chemistry, soil sciences and physiology (whose terms have no value connotations). Then, 'using X is safe' may be concluded,[ix] – usually qualified by 'provided that it is used in accordance with stipulated regulations' – if the investigations confirm that none of the investigated effects would occur with significant magnitude and probability when X is used in accordance with the regulations. This account is consistent with the picture of scientific safety studies that has step (1) preceding steps (2) and (3); but it clarifies that the move from empirically confirmed results at (1) to the claim the value-implicated 'X is safe' and to value judgments of relevance at (2) rests upon value judgments made at step (0). It follows that the conclusion, 'X is safe', might appropriately be challenged – without thereby challenging the scientists' judgments about each of the particular possible effects investigated – on the basis of the value judgment that not all the harmful possible effects of using X were identified at (0).

The outcomes of "scientific" safety studies usually constitute the only input to the deliberations of the 'technical' commissions that participate in public policy deliberations about using and regulating technoscientific objects. In these studies (in the GEO case), at step (0), the possible effects identified as harmful are a subset of those that may be occasioned by way of physical/chemical/biological mechanisms directly triggered by events within the modified genomes of plants. One can identify *two ways in which the adequacy of these studies might be challenged*.[x]

First: Conclusions drawn about the safety of using V (a genetically engineered plant variety) could be challenged on the ground that the subset chosen for investigation does not include some possible effects, with similar mechanisms, that are of of special salience for those who uphold a particular value-outlook.[xi] For them, even well conducted studies on the the items of the subset chosen will be insufficient to confirm that using V is safe.[xii] Challenges of this type can be

resolved (in principle) by conducting more scientific studies of the same kind after having identified a larger relevant subset.[xiii]

Second: Their adequacy could be challenged by those, who object that the set from which the subsets are chosen for "scientific safety studies" is not sufficiently encompassing. For them, deliberations about the safety of using GE-plants should be informed by appropriate empirical investigations, not only of potential effects occasioned by way of physical/chemical/biological mechanisms directly triggered by events within their modified genomes, but also the full range of potential ecological and socioeconomic effects occasioned by using them in the environments (agroecosystems) of their actual or intended use, and under the socioeconomic conditions of their use, taking fully into account that the potential effects vary from variety to variety and species to plant species. Upholding values of respect for human rights, democratic participation and environmental sustainability, which are opposed to those of capital and the market, often motivates challenges of this kind. These potential effects cannot *all* be investigated in "scientific safety studies," for they require utilizing ecological, human and social categories that have no place in research in such areas as physics, chemistry, and molecular biology, and that may include thick ethical terms (e.g., food security, being poisoned).[xiv] To investigate them empirically, therefore, requires adopting methodological approaches that are not reducible to those used in the indicated scientific areas, and that are generally outside of the expertise of scientists trained in the methodologies appropriate to them. The expertise required to engage in research that leads to the development of GEOs is quite different from that required for studies about the safety of using them.

At issue here are not only concerns about risks (potential harmful effects). Farmers (and their communities) in many areas of the world have suffered serious health problems because of having been exposed to glyphosate (the principal active ingredient in the widely used herbicide, RoundUp) sprayed on fields planted with glyphosate-resistant GEOs.[xv] They are unimpressed when told that the varieties of GEOs planted in these fields had undergone and passed "scientific safety tests." They know from their experience (even if it is not well recorded in peer reviewed studies) that, regardless of what was the case in the conditions of the tests, it is not safe to cultivate these GEOs (which require the accompanying use of glyphosate) in the ways and under the conditions in which they are used in their locales. And, they continue to be unimpressed when the manufactures and regulators of the GEOs insist that the problem was not with cultivating the GEOs, but with using glyphosate without heed to stipulated regulations for safe use,[xvi] for they have good reason to believe that the sellers of GEOs and glyphosate know that they will in fact not be used in accordance with these regulations.[xvii]

3.

Summing up, ethical and social values properly affect decisions (at step 0)) made about the criteria to be deployed for identifying the range of risks that should be considered in public policy deliberations, and of the relevant kinds of empirical data needed for making judgments about them. They also – consistent with maintaining that judgments about safety (step (1)) can be settled

prior to steps (2) and (3) – also affect the standards deployed for determining how well claims about risks should be supported (by the available empirical data) – in order to ensure that risks are dealt with properly in public policy deliberations.

Those who uphold values of capital and the market (agribusiness corporations, governments that prioritize economic growth, etc) are likely to cite the ethical gravity of losses (both economic and, allegedly, for food security) that would be incurred by failing to use GEOs on a wide scale; and consequently to require less stringent standards of evidential appraisal than those who uphold values of respect for human rights, democratic participation and environmental sustainability, who are likely to adopt precautionary stances that permit time for research incorporating more stringent standards to be met.[xviii] Similarly, those who uphold the latter values are likely to emphasize the importance of step (3): investigating alternatives to the food/agricultural system, in which using GEOs and the use of agrotoxics are acquiring ever larger roles, alternatives such as agroecology, a scientifically-informed approach to agriculture that attends simultaneously to production, sustainability, social health, strengthening the values and cultures of local communities, and to furthering the practices needed to implement policies of food sovereignty – and to urge the public support of research, in which are adopted strategies appropriate for dealing with the human, ecological and social dimensions of agroecosystems.[xix]

Scientists, *qua* scientists, however, do not have authority in the realm of ethical and social values. The values they uphold, even when widely shared, do not trump those upheld by other groups in democratic public policy deliberations. Thus, their judgments, about the evidential support that claims about the safety of planting GEO crops and consuming their products have, may sometimes be reasonably contested partly on value-laden grounds (cf. de Melo-Martín & Intemann, 2017, p. 131). That contestation cannot be rebutted by appeal to the alleged "scientific consensus" that GEOs (or, particular varieties of them) are safe. Apart from the fact that actually there is no such consensus, manifestly so among experts in biosafety investigations,[xx] if there were, it would likely secrete the scientists' shared value commitments, a matter on which they have no authority. Appeal to such an alleged consensus covers up the role of upholding the values of capital and the market in affirming it.

It follows that, in the context of deliberations about public policy, the trustworthiness of scientists is put into question and their authority unmerited,

- unless they engage with representatives of all stakeholders in the outcomes of the policies (as, for the most part, has not happened in the GEO case);
- unless, moreover, in doing so – respecting what Longino (2002, p. 129–135) calls "tempered equality" of participants in the deliberations – , they take into account that upholding competing values (e.g., of company-employed scientists and family farmers) may lead to making different judgments concerning relevant data, hypotheses to investigate, and approaches to farming, as well as concerning the degree of support required for claims about safety to merit credibility.

4.

5

Let us now return to the three claims (introduced at the outset) that the 135 Nobel laureates endorsed:[xxi]

These claims are ambiguous, misleading, in some instances false, and apparently made without acquaintance with the relevant studies and arguments of their critics. (i) is false: I am not aware of any agency that has compared the safety of GEO crops and their food products with that of agroecological (or organic farming) methods of production – the agencies have not sought out the results of research dealing with that comparison (and very little of it has been conducted). At most, they have found GEO crops and products to be at least as safe as conventional high-input crops and their products, but that doesn't respond to the critics who endorse agroecological methods of production. (ii) is probably true – but misleading: it does not mention that epidemiological studies of consumption of GEOs have not been conducted,[xxii] to a large extent because legal prohibition of labelling GEO products poses probably an insurmountable impediment to conducting them; and that it is well documented that cultivating GEOs has occasioned health problems for numerous farmers who have been exposed to the agrotoxics, whose use is integral to the cultivation of certain varieties of GEOs. (iii) is ambiguous: the environmental impacts may indeed be less damaging than those of conventional high-input agriculture; but they are incomparably more damaging to the environment than agroecological farming that has environmental sustainability built into its fundamental objectives.

By dismissing criticisms like these "based on emotion and dogma contradicted by data," and not attempting to rebut them in a context where something like Longino's conditions are in place, the scientists undermine the authority that science should be able to demand to be recognized; and they weaken the contribution that science could make to democratic policy deliberations.

References

Bombardi, Larissa M. (2017) Geografia do Uso de Agrotóxicos no Brasil e Conexões com a União Europeia. E-book, <https://drive.google.com/file/d/1ci7nzJPm_J6XYNkdv_rt-nbFmOETH80G/view>. São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

De Melo-Martín, I. and Intemann, K. (2018) *The Fight against Doubt: How to bridge the gap between scientists and the Public*. New York: Oxford University Press.

Hilbeck, A., Binimelis, R., Defarge, N., Steinbrecher, R., Székács, A., Wickson, F., Antoniou, M., Bereano, P. L., Clark, E. A., Hansen, M., Novotny, E., Heinemann, J., Meyer, H., Shiva, V. & Wynne, B. (2015) No scientific consensus on GEO safety. *Environmental Sciences Europe* 27: 4–9.

Human Rights Watch (2018) The Failing Response to pesticide Drift in Brazil's Rural Communities, July 20, 2018, <https://www.hrw.org/report/2018/07/20/you-dont-want-breathe-poison-anymore/failing-response-pesticide-drift-brazils>.

Krimsky, S. (2015) An illusory consensus behind GEO health assessment. *Science, Technology and Human Values* 40 (6): 883–914.

Lacey, H. (2005) *Values and Objectivity in Science; current controversy about transgenic crops.* Lanham, MD: Lexington Books.

— (2015a) Food and agricultural systems for the future: science, emancipation and human flourishing. *Journal of Critical Realism* 14 (3), 2015: 272–286.

— (2015b) Agroécologie : la science et les valeurs de la justice sociale, de la démocratie et de la durabilité. *Ecologie et Politique*, No. 51, 2015: 27–40.

— (2016) Science, respect for nature, and human well-being: democratic values and the responsibilities of scientists today. *Foundations of Science* 21(1): 883–914.

— (2017) The safety of using genetically engineered organism: empirical evidence and value judgments. *Public Affairs Quarterly* 31 (4): 259–279.

Lacey, H., Corrêa Leite, J., Oliveira, M.B., & Mariconda, P.r. (2015a) Transgênicos: malefícios, invasões e diálogo. *JC Notícias*, Edition 5167 (April 30, /2015), http://www.jornaldaciencia.org.br/edicoes/?url=http://jcnoticias.jornaldaciencia.org.br/9-transgenicos-maleficios-invasoes-e-dialogo/.

— (2015b) Transgênicos: diálogo. *JC Notícias*, Edition 5182 (May 22/2015), http://www.jornaldaciencia.org.br/edicoes/?url=http://jcnoticias.jornaldaciencia.org.br/27-transgenicos-dialogo/.

Longino, H. (2002) *The Fate of Knowledge*. Princeton: Princeton University Press.

Laureates Letter (2016) "Laureates letter supporting precision agriculture," http://supportprecisionagriculture.org/nobel-laureate-gmo-letter_rjr.html .

US National Academies of Science, Engineering and Medicine (2017). *Genetically Engineered Crops: Experiences and Prospects*. Washington: National Academies Press.

Paganelli, A., Gnazzo, V, Acosta, H., López, S.L. & Carrasco, A.E. (2010) 'Glyphosate-based herbicides produce teratogenic effects on vertebrates by impairing retinoic acid signaling'. *Chemical Research in Toxicology* 23: 1586–1595.

Traavik, T. & Ching, L.L. (2007) *Biosafety first: Holistic approaches to risk and uncertainty in genetic engineering and genetically modified organisms*. Trondheim, Norway: Tapir Academic Press.

Appendix

The central concern of the letter signed by the Nobel laureates is to support the program of research on Golden Rice [a variety of genetically engineered rice] and to denounce opposition to it, especially that of the NGO, Greenpeace. In a longer work, I would also discuss critically the way in which the letter misleads both about the state of research on Golden Rice and about that character of criticisms that question the importance of this research.

(a) The letter states that Greenpeace "has spearheaded opposition to Golden Rice, which has the potential to reduce or eliminate much of the death and disease caused by a vitamin A deficiency, which has the greatest impact on the poorest people in Africa and Southeast Asia". It called upon "governments of the world to reject Greenpeace's campaign against Golden Rice specifically, and crops and foods improved through biotechnology in general; and to do everything in their power to oppose Greenpeace's actions and accelerate the access of farmers to all the tools of modern biology, especially seeds improved through biotechnology"; and concluded with the warning: "Opposition based on emotion and dogma contradicted by data must be stopped," accompanied by the rhetorical question: "How many poor people in the world must die before we consider this a 'crime against humanity'?"

(b) Around the same time, the US National Academies of Science, Engineering and Medicine (2017) pointed out that the International Rice Research Institute (IRRI) had stated reported: "Golden Rice will only be made available broadly to farmers and consumers if it is successfully developed into rice varieties suitable for Asia, approved by national regulators, and shown to improve vitamin A status in community conditions. If Golden Rice is found to be safe and efficacious, a sustainable delivery program will ensure that Golden Rice is acceptable and accessible to those most in need" (p. 228). As of July 2016, IRRI was continuing research on developing varieties of Golden Rice for use in SE Asia, and (according to it) none of the conditions it stated had yet been met - it is for this reason that Golden Rice has not been introduced.

(c) Two years later, earlier this year (2018), IRRI asked the USFDA for an opinion regarding the safety of a variety of Golden Rice (called GR2E - the only variety yet submitted for regulatory approval - but not yet approved in any Asian country). FDA (May 24, 2018) endorsed the evaluation of IRRI (and the Australian regulatory body) that GR2E is safe for consumption, while pointing out that it is not intended for food or animal uses in USA. However, it added: "the concentration Beta-carotene in GR2E rice is too low to warrant a nutrient content claim." GR2E is safe but not nutritionally relevant.

(d) The signers of the letter, thus, were remarkably uninformed about the state of research on Golden Rice – and also about the views and stances of Greenpeace (I am not associated with Greenpeace). On its website Greenpeace states that its objective is to "ensure the ability of Earth to nurture life in all its diversity." It fits into the body of critics of using GMOs, who maintain that the dominant food-agricultural system (in which using GEOs has become for the time being a fundamental component) cannot respond adequately to the food and nutrition needs of the world's impoverished peoples (and the right to food security for everyone), and that these needs can best be ameliorated by the programs of agroecology and food sovereignty (Lacey, 2015a; 2015b) – and that programs for developing GEOs (like Golden Rice) are taking resources away from developing effective and lasting solutions to death and disease caused by vitamin A deficiency. Greenpeace has a respected place among these critics (and its "direct actions" and contributions to legal challenges are often appreciated by them). Of course, it would be legitimate to rebut the critics with argument and evidence. One wonders why the laureates did not attempt to do so.

(e) The credibility of pronouncements made by scientists of outstanding achievement is weakened when they sign letters like this one, accompanied by inflated, emotionally charged rhetoric, that has has a slender basis in fact. It would be enhanced if they entered into the type of dialogue, advocated by Helen Longino, in which scientists would "listen to" the evidence provided by relevant parties, attempt to understand critics, and not tar them without a hearing. Science has an indispensable contribution to make in policy deliberations; but it is not the determiner of policy. Science will be enhanced, and its role in democratic societies consolidated, if it claims only to have authority where it is actually warranted.

Notes

i **DRAFT** (not for citation outside of the PSA meeting in Seattle) – October 15, 2018. The text is a draft of the presentation I'm planning to make. The notes contain details that will be incorporated into an eventual completed paper.

ii See Appendix.

iii E.g., Mark Lynas (Cornell Alliance for Science), *A plea to Greenpeace*, <http://www.marklynas.org/2016/06/a-plea-to-greenpeace/>.
In this paper I only consider GEOs used in agriculture. I take for granted that claims to the effect that using GEOs is safe refer to GEOs that have passed safety tests, including those currently available on the market. (Obviously an unsafe GEO could be developed. Some varieties of GEOs have been developed that, after failing to pass safety tests, were not released for use.)

iv More fully developed and defended in Lacey (2005), Part 2.

v Deliberations concerning (2) and (3) cannot be settled in scientific inquiry (sound empirical inquiry), but there are sound empirically-based inputs that are (or could be) relevant to them. The deliberations will not be satisfactory if they do not draw upon these inputs. (See Lacey, 2005.)

vi Claims about efficacy need to be stated in a more qualified and nuanced way. I also will not contest that the claim that scientific research has not provided compelling evidence that consuming GEO products is unsafe health-wise. (The absence of compelling evidence that GEO products are unsafe to consume does not mean that there is compelling evidence that they are safe to consume – it depends on whether or not the necessary research has been conducted.)

vii The *ceteris paribus* qualification is needed to take into account that sometimes considerations, not reducible to safety ones, may properly be appealed to.

viii I will not discuss here how this set is generated – e.g., from considering past investigations, role of values in it, stakeholders' concerns, etc) – and who (holding what values?) makes (and should make) the identification of what should be considered harmful? following what kinds of deliberations? and who should be represented in the deliberations?.

ix To conclude on the basis of empirical investigation that 'X is safe' requires showing one-by-one that each member of the set of anticipated effect (judged to be harmful) is unlikely to occur at sufficient magnitude under the conditions imposed by proposed regulations. This presupposes: (a) an inductive move to unanticipated effects; and (b) that representative cases of all the effects, that should be labelled potentially harmful, are members of the set.

x I have argued elsewhere that here methodological and value considerations mutually reinforce each other (Lacey, 2017). Proponents of using GEOs often say that these safety studies investigate the risks occasioned by the GEOs themselves, and not those occasioned by the accompaniments of using them in agroecosystems or by socioeconomic mechanisms.

xi E.g., effects on soil microorganisms, a matter especially salient for those who regard maintaining soil fertility as indispensable for sustainable agriculture.

xii The studies, which have produced many of the results that have actually informed public policy and regulatory decisions, have been criticized for having a number of kinds of shortcomings (e.g., connected with conflicts of interest, and the use of intellectual property rights to maintain studies secret and so unavailable for replication and independent confirmation). Value judgments pervade these criticisms and their rebuttals. I will not attend to the questions that arise here.

xiii Such challenges might be deemed irrelevant by those who reject the value-outlook for which the possible effects have special salience, and so who reject the need for the further studies. Those adhering to the values of capital and the market sometimes take such a stand. How reasonable that might be depends on the arguments offered against holding the value-outlook in question.

xiv For elaboration see Lacey (2016; 2017).

xv For documentation, see, e.g., Bombardi (2017); Paganelli, et al. (2010); Human Rights Watch (2018).

xvi After a jury in California recently ruled that Monsanto was responsible for a man's being afflicted with cancer, and imposed a huge fine on it because it – for it was deemed that Monsanto had "acted with malice" in not providing warning on its label of the risks to health occasioned by using Roundup – the President of Bayer (that has now incorporated Monsanto) responded: "The correct use of Roundup doesn't present a risk to health" (reference to be added). [Monsanto has appealed the ruling.]

xvii Three years ago, when representatives of farmers – who had been poisoned in this way – came to present their testimony at a meeting of the "technical" commission in Brazil (CTNBio) that had appraised a particular variety of GEOs as safe, they were not granted a hearing since (most members of the commission maintained) they were bearers only of anecdotal (not scientific) evidence that had no relevance to the conclusions of scientific safety studies. When they then disrupted the meeting (and others of their group prevented the planting of a new variety of GEOs by invading a nursery and pulling up all the seedlings), they were denounced by major scientific organizations as having no respect for science, and acting on the

basis of "emotion and dogma." For criticisms of this stance taken by the majority of members of CTNBio, and a response to a rebuttal of the criticism, see Lacey, et al. (2015a; 2015b), articles published in *JC Notícias*, a daily e-newsletter of *Jornal da Ciência*, a publication of SBPC (Brazilian Society for the Advancement of Science).

The narrow scope of "scientific safety studies" is sometimes justified on the ground that the investigations of the social impact of using GEOs is not "scientific," for the methodologies adopted in them are not reducible to those adopted in the mainstream areas of science mentioned above. Be that as it may: I won't quibble about how to use the term "scientific" (a thick ethical term); the investigations in question are (when properly conducted) systematic empirical investigations. If they don't count as "scientific", that would imply that the results of "scientific" investigations cannot provide sufficient input into deliberations concerning public policies about safety, and would need to be supplemented with input from other kinds of empirical investigations.

[xviii] See Lacey (2017).

[xix] For details, see Lacey (2005; 2015a; 2015b).

[xx] See, e.g., Hilbeck, et al. (2015); Krimsky (2015); Traavik & Ching (2007).

[xxi] See Appendix.

[xxii] Unless all the relevant research has been conducted (and it has not been in this case), the absence of compelling evidence that GEO products are unsafe to consume does not imply that there is compelling evidence that they are safe to consume – and it has nothing to do with harms that may be caused by, e.g., contact with an agrotoxic, rather than by consumption.

The Reference Class Problem for Credit Valuation in Science

Carole J. Lee (c3@uw.edu)

Abstract: Scholars belong to multiple communities of credit simultaneously. When these communities disagree about how much credit to assign to a scholarly achievement, this raises a puzzle for decision theory models of credit-seeking in science. The reference class problem for credit valuation in science is the problem of determining to which of an agent's communities – which reference class – credit determinations should be indexed for any given act under any given state of nature. I will identify strategies and desiderata for resolving ambiguity in credit valuation due to this problem and explain how pursuing its solution could, ironically, lead to its dissolution.

1. Introduction

Within the scientific community, there is a common understanding that its reward system drives problematic behavior linked to publication patterns, pipeline retention, hypercompetitive scientific cultures, and reproducibility. Conversely, there is also a shared sentiment that, in order to change these cultures and behaviors in ways that would improve science, the scientific community must coordinate across institutions to change how credit is assigned at the level of the individual scientist (Alberts et al. 2014, Nosek et al. 2015, Aalbersberg et al. 2017, National Academies of Sciences 2018, National Science Foundation 2015, Blank et al. 2017). The hope is

that increasing individual researchers' incentives towards increased transparency and openness will improve the integrity, reproducibility, and accuracy of the published record.[1]

Analogously, philosophers working in the "credit economy" tradition adopt the working assumption that there is some amount of credit that agents can accrue for different acts under different states of nature. This assumption allows them to use decision theory to model how credit-seeking among individual scientists can give rise to behavior and norms that support or thwart the achievement of community-wide goals. When, in the aggregate, individual credit-seeking cuts against collective ends, their approach can explore how changes to individuals' incentive structures can nudge and redirect individual behavior (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Bright 2017, Heesen 2017, Kitcher 1990, Strevens 2003, Zollman 2018). Different philosophers make different assumptions about the norms by which credit gets allotted – for example, whether credit is best thought of as all-or-nothing (Strevens 2003, Bright 2017, Heesen 2017) or as something that may come in degrees (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Zollman 2018). However, the general approach assumes that there is some precise way to assign credit to different acts under different states of nature – an assumption that allows these philosophers to model credit-seeking behavior and the emergence of scientific norms in formally tractable ways.

But, how much credit gets assigned to any given act under any given state of nature? Just as each of us simultaneously belongs to multiple social categories each of which is tied to implied social hierarchies (Macrae, Bodenhausen, and Milne 1995, Crenshaw 1989), each

[1] Institutions can also experience incentives that promote or thwart scientific ends (Lee and Moher 2017).

scholar simultaneously belongs to multiple communities of value with implied social hierarchies for assigning credit. To which of an agent's communities – which reference class – should credit determinations be indexed and why?

In this paper, I will use examples from the current context of science's complex and dynamic culture to motivate and illuminate what I will call the *reference class problem for credit valuation in science*. I will identify a few strategies and desiderata for solving ambiguity in credit assignments due to the reference class problem. And, I will say a bit about how developing the resources needed to solve it could ultimately sow the seeds for its own dissolution.

2. The Reference Class Problem for Credit Valuation in Science

The contours of this puzzle about the "coin of recognition" (Merton 1968, 56) become visible when one moves beyond thinking about credit in generic, abstractions of scientific communities towards the heterogeneous communities we find today. I start from this slightly more concrete perspective because prestige requires recognition *by individuals and forums* that are themselves valued by credit-seeking scholars (Zuckerman and Merton 1971, Lee 2013): credit worthiness in science is a function of the individuals and systems designed to assess, allocate, dispute, and enforce it. Although some aspects of Zuckerman and Merton's narrative about the origins of the normative structure of science have been contested by historians (Csiszar 2015, Biagioli 2002), we see the social dynamics Zuckerman and Merton proposed clearly at play in contemporary science. For example, Nature Publishing Group recently found that – for the 18,354 authors in science, engineering, and medicine surveyed – the reputation of a journal is the primary factor driving choices about where to submit their work, where reputation is

primarily determined by the journal's impact factor and whether it is "seen as the place to publish the best research" (Nature Publishing Group 2015). Factors associated with a journal's ability to archive and disseminate research – things like a journal's time from acceptance to publication, indexing services, or Open Access options – were much less important.[2]

Within academia, each of us simultaneously belongs to multiple communities of value. The reference class problem arises when these different communities of value disagree about the amount of credit an agent accrues for choosing some act under some state of nature. Although I take this problem is be general, for the sake of clarity and simplicity in presentation, I will focus my examples on communities that can be described as having a nesting structure: for example, individual scholars belong to specific sub-disciplines, which are nested within disciplines, which are nested within a more general population of scholars. A sub-population that is nested within a population can have a credit sub-culture whose valuations differ from that of the population, whose valuations can differ from that of the super-population. In these cases, changing how narrowly or broadly one draws the boundaries of an agent's community of valuation can change the amount of credit assigned to a scholarly accomplishment. This gives rise to the *reference class problem for credit valuation in science*: to which of the agent's communities – which reference class – should credit valuations be indexed when determining the amount of credit the agent accrues for different acts under different states of nature?

[2] I recognize that some decision theorists, especially those working outside of philosophy, may reject or remain agnostic about attributing mental states such as beliefs to agents (Okasha 2016). However, because I understand credit and credit-seeking as sociological phenomena involving status beliefs such as these, I am committed to attributing beliefs to agents.

There are many examples across academia where nesting community structures can give rise to paradoxes and pathologies in credit assignments. For example, scholars' individual sense of what counts as quality work – their individual credit assignments – may deviate from what is endorsed in a sub-discipline or discipline's status hierarchy (Correll et al. 2017, Centola, Willer, and Macy 2005, Willer, Kuwabara, and Macy 2009). A puzzle that has cachet in a sub-discipline may be of peripheral importance within that discipline: for example, a more accurate technique for measuring how temperature cools with elevation considered critical in mountain meteorology and mountain ecology (Mindner, Mote, and Lundquist 2010) may have less visibility, despite its relevance, to the larger discipline of hydrology (Livneh et al. 2013). A question or technique that is thought to have high impact across fields (e.g., machine learning) may have little prominence within some of those fields.

Hypothetically speaking, one could imagine differences in valuations giving rise to a *Simpson's paradox in credit valuation*. Simpson's paradox is a phenomenon whereby a trend that appears in a population reverses or disappears when it is disaggregated into sub-populations (Blyth 1972). For example, a classic study found that, when looking at aggregate graduate school admissions data at UC Berkeley, women were, on the whole, less likely than men to be accepted; however, when the data was disaggregated into admitting departments, women were more likely than men to be admitted (Bickel, Hammel, and O'Connell 1975). Analogously, a *Simpson's paradox in credit valuation in science* would occur in cases where a population-level preference for scholarly product *a* versus *b* reverses when the population is disaggregated into its component sub-populations. In Simpson's Paradox cases, thinking more carefully about the context of evaluation usually leads to using a reference class that is finer-grained than the population-level. However, it's not clear whether this would always be the case in evaluations of

scientific credit. Hypothetically speaking, consider a hypothetical scenario in which an interdisciplinary project is not preferred by the individual disciplines represented by its authors or content, but is preferred when those disciplines are aggregated together. And, imagine that this project gets published in a journal, valued by those disciplines, that seeks papers of interest *across and beyond disciplines* (not just within disciplines): this is one way to interpret, for example, *Science's* mission to publish papers that "merit recognition by the wider scientific community and general public. . . beyond that provided by specialty journals" (Science). Which reference class would be most relevant in evaluating the value of this project?

There are other ways of dividing scholarly communities into nesting structures that create tensions in credit assignments. The pressures a scholar may feel from the incentive structure impacting her department/school may be slightly different from the incentive structure impacting her university. A coarse but concrete way to see this is to think about the prestige structure reified and reinforced by ranking systems (Espeland and Sauder 2012, 2016, Sauder and Espeland 2006), which transform "the ways professional opportunities are distributed" (Espeland and Sauder 2016, 7). An untenured business school professor with a potentially high impact manuscript needs to burnish her prestige in the eyes of both her dean and her provost, since both will evaluate her tenure case. If her provost is working to gain stature on the Academic Rankings of World Universities [ARWU], the professor should submit her manuscript to *Science* or *Nature*, since the ARWU ranks universities by their publications in these journals (Academic Ranking of World Universities 2018). However, if her dean is trying to gain stature on the *Financial Times* International ranking of MBA programs, she should submit to one of the fifty business, economics, or psychology journals by which the FT ranking system evaluates Business

school prestige – notably, the journal list does not include *Science* or *Nature* (Ormans 2016). What should the business school professor do?

Finally, credit assignments can vary depending on how long a time window a scholar keeps in view. A coarse but concrete way to think about this is by looking at how metrics for evaluating scholarship change over time. Journal impact factors are becoming less useful measures for evaluating an individual's scholarly contribution: since the advent of the digital age, the most elite journals (including *Science* and *Nature*) are publishing a decreasing percentage of the top cited papers (Larivière, Lozano, and Gingras 2013); the relationship between journal impact factor and paper citations has declined over time (Lozano, Larivière, and Gingras 2012); and, the citation distributions between journals "overlap extensively" (Larivière et al. 2016). The current wisdom is that if quantitative indicators are to be used to evaluate research, it is more useful to use article-level metrics such as citations as well as alternative metrics such as downloads and views (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017). On the horizon, there are now calls for creating new metrics that can encourage researchers and journals to be transparent and open in their reporting practices (National Academies of Sciences 2018, Wilsdon et al. 2017, Aalbersberg et al. 2017). Note that, the rise of such metrics – as well as the growing meta-research literature that ranks journals by the replicability (Schimmack 2015) or sample size and statistical power of their published results (Fraley and Vazire 2014) – makes it possible for a journal's impact factor and epistemic credibility to come apart (Fang and Casadevall 2011).

Decision theorists capture the risky nature of individual choices by allowing for uncertainty about which states of the world will come to be; and, when the probabilities attached to different outcomes are understood subjectively, these models permit a kind of subjectivity in

estimates of expected credit for different acts. However, I hope the examples throughout this section animate genuine *ambiguity in credit* due to the reference class problem for credit valuation in science.

3. Strategies and Desiderata for Solving the Reference Class Problem

How might decision theorists try to solve the reference class problem for assigning credit in science? One possible approach argues for the "correctness" of using one community rather than another. For example, it might be tempting to argue that all prestige is discipline-based since many scholarly prizes are distributed for excellence in particular disciplines (e.g., Nobel prize, Fields prize, academic society prizes); and, even when research is funded or published in interdisciplinary contexts, it may be primarily evaluated on the basis of its disciplinary excellence (Lamont 2009, but see Lee et al. 2013). Indexing credit valuation to a particular community need not prevent scholars from outside that community from understanding the relative value of that contribution: for example, if one were to adopt the old-fashioned and problematic assumption that an article's impact can be measured by the impact factor of the journal in which it is published,[3] and one recognizes that citations rates vary across disciplines, one could use field-normalized percentiles to understand a paper's impact in a metric that is legible across fields (Hicks and Wouters 2015). Because this strategy for addressing the

[3] The citation distributions within journals are so skewed that it is statistically improper to infer the impact of an individual article on the basis of the impact factor of the journal in which it is published (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017, Larivière et al. 2016, Wilsdon et al. 2015).

reference class problem relies heavily on identifying the "right" community, defending the centrality of the chosen community as opposed to others is critical. For example, some may challenge the idea that disciplines should be the sole arbiter of credit: note that the awarding of some scientific prizes reach across disciplinary conceptions of excellence (e.g., consider winners of the MacArthur Genius Prize and the psychologists who have won the Nobel Prize in Economics).

Another possible approach creates an algorithm that calculates the credit value of a scholarly contribution by summing the credit valuation of multiple communities. This approach would need to identify exactly how much to weight each community's valuation – with a rationale for why – since different weightings could lead to different overall credit valuations.[4] Note that some scholars take this style of approach when trying to measure the relative prestige of journals: in particular, the Eigenfactor score rates journals according to the number of its incoming citations, where the "relative importance" of each incoming citation is contextualized by the frequency with which the citing journal is itself cited (West, Bergstrom, and Bergstrom 2010).

Those who may wish to model the implications of different approaches for solving the reference class problem may try to do so by setting up hypothetical communities that assign

[4] On the face of it, this may seem like a form of commensuration because it involves summing values to calculate an overall score (Espeland and Stevens 1998). However, the process of commensuration requires combining values across *qualitatively* different domains of value. For clearer examples of commensuration in scholarly evaluation, see Lee (2015).

community boundaries and credit assignments in *de facto* ways to see what kinds of behaviors and norms emerge.

However, to solve the underlying conceptual problem, one must provide theories of community and credit that address two fundamental but vexing questions. How should one define and gerrymander the boundaries of the relevant communities invoked in the proposed solution? And, how does one determine the amount of credit those communities would assign to different acts under different states of nature? These questions may not be independently answerable. The boundaries of a community may need to be defined in terms of patterns of shared lore among its members about how credit is accrued – shared beliefs that coordinate credit-seeking and enforcement behavior in cases where status beliefs are internalized as norms (Merton 1973) and in cases where they are not (Willer, Kuwabara, and Macy 2009, Ridgeway and Correll 2006). Conversely, in recognition that some community members can have more influence than others on the content of reigning status beliefs, a community's credit assignments may need to be defined with some reference to the causal patterns of interaction among specific individuals and clusters of individuals – including status judges who wield "social control through their evaluation of role-performance and their allocation of rewards for that performance" (Zuckerman and Merton 1971, 66). Note, however, that answers to these questions should not *exclusively* inform each other. Notably, we must be careful not allow the size of a scholarly population and/or the power of its status judges to fully determine the intellectual value of the questions pursued by any particular partition of the scholarly universe.

4. Conclusion

Scientific credit – the "coin of recognition" (Merton 1968, 56) – is assessed, allocated, disputed, and enforced by many different communities and institutions within science that support and sustain a multiplicity of status hierarchies. This gives rise to what I have called the reference class problem for credit valuation in science. Solving this problem requires developing rich theories of community and credit that are based on fine-grained information about the structure and status systems of complex scholarly networks. The irony of this assessment is that such investigation towards solving the reference class problem could ultimately sow the seeds for its own dissolution.

In particular, such study can render friable a critical assumption for both the reference class problem and for decision theory models: namely, that communities, once defined, assign determinate amounts of monistic credit for different acts under different states of nature – that credit "can vary quantitatively but not qualitatively" (Anderson 1993, xii).[5] Contrary to this, recent policy papers call for moving away from narrowly conceived measurements of research excellence towards broader ones that are sensitive to the diversity of individual researchers', programs', and academic institutions' research missions (Hicks and Wouters 2015, Wilsdon et al. 2015). Such work can include community-engaged scholarship that creates, disseminates, and implements knowledge in coordination with the public to identify social interventions, change social practice, and influence policy (Hicks and Wouters 2015, San Francisco Declaration on Research Assessment 2013, Boyer 1990, Escrigas et al. 2014). From the

[5] Note too that, for formal reasons, the assumption that individual credit assessments could be aggregated into a collective one is questionable given the challenges of combining individual preferences into collective ones (Arrow 1950).

perspective of these efforts, plurality in our notions of scholarly excellence and credit – and differences in valuation and prioritization practices between individuals and communities – may be best conceived, not as a logical problem to solve, but as a starting point for theorizing.

Acknowledgments: Many thanks to Christopher Adolph, Aileen Fyfe, Crystal Hall, Jessica Lundquist, Conor Mayo-Wilson, and Kevin Zollman for helpful conversations. This research used statistical consulting resources provided by the Center for Statistics and the Social Sciences, University of Washington.

References

Aalbersberg, IJsbrand Jan, Tom Appleyard, Sarah Brookhart, Todd Carpenter, Michael Clarke, Stephen Curry, Josh Dahl, Alex DeHaven, Eric Eich, Maryrose Franko, Len Freedman, Chris Graf, Sean Grant, Brooks Hanson, Heather Joseph, Véronique Kiermer, Bianca Kramer, Alan Kraut, Roshan Kumar Karn, Carole Lee, Aki MacFarlane, Maryann Martone, Evan Mayo-Wilson, Marcia McNutt, Meredith McPhail, David Mellor, David Moher, Alison Mudditt Mudditt, Brian Nosek, Belinda Orland, Tim Parker, Mark Parsons, Mark Patterson, Solange Santos, Carolyn Shore, Dan Simons, Bobbie Spellman, Jeff Spies, Matt Spitzer, Victoria Stodden, Sowmya Swaminathan, Deborah Sweet, Anne Tsui, and Simine Vazire. 2017. "Making science transparent by default; Introducing the TOP Statement." *OSF Preprints*. doi: https://doi.org/10.31219/osf.io/sm78t.

Academic Ranking of World Universities. 2018. "ShanghaiRanking's Academic Ranking of World Universities 2018 Press Release." accessed September 1.

http://www.shanghairanking.com/Academic-Ranking-of-World-Universities-2018-Press-Release.html.

Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. 2014. "Rescuing US biomedical research from its systematic flaws." *Proceedings of the National Academy of Sciences* 111 (16):5773-7.

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Arrow, Kenneth J. 1950. "A difficulty in the concept of social welfare." *Journal of Political Economy* 58 (4):328-46.

Biagioli, Mario. 2002. "From Book Censorship to Academic Peer Review." *Emergences: Journal for the Study of Media & Composite Cultures* 12 (1):11-45.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. "Sex bias in graduate admissions: Data from Berkeley." *Science* 187 (4175):398-404.

Blank, Rebecca, Ronald J. Daniels, Gary Gilliland, Amy Gutmann, Samuel Hawgood, Freeman A. Hrabowski, Martha E. Pollack, Vincent Price, L. Rafael Reif, and Mark S. Schlissel. 2017. "A new data effort to inform career choices in biomedicine." *Science* 358 (6369):1388-9.

Blyth, Colin R. 1972. "On Simpson's Paradox and the sure-thing principle." *Journal of the American Statistical Association* 67 (338):364-66.

Boyer, Ernest L. 1990. *Scholarship Reconsidered*. San Francisco, CA: The Carnegie Foundation for the Advancement of Teaching.

Bright, Liam Kofi. 2017. "On Fraud." *Philosophical Studies* 174:291-310.

Bruner, Justin, and Cailin O'Connor. 2017. "Power, Bargaining, and Collaboration." In *Scientific Collaboration and Collective Knowledge*, edited by Thomas Boyer-Kassem, Conor Mayo-Wilson and Michael Weisberg, 135-157. Oxford, UK: Oxford University Press.

Centola, Damon, Robb Willer, and Michael Macy. 2005. "The emporer's dilemma: A computational model of self-enforcing norms." *American Journal of Sociology* 110 (4):1009-40.

Correll, Shelley J., Cecilia L. Ridgeway, Ezra W. Zuckerman, Sharon Jank, Sara Jordan-Bloch, and Sandra Nakagawa. 2017. "It's the conventional thought that counts: How third-order inference produces status advantage." *American Sociological Review* 82 (2):297-327.

Crenshaw, Kimberle. 1989. "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics." *University of Chicago Legal Forum* 139:139-168.

Csiszar, Alex. 2015. "Objectivities in Print." In *Objectivity in Science: New Perspectives from Science and Technology Studies*, edited by Flavia Padovani, Alan Richardson and Jonathan Y. Tsou, 145-69. Cham, Switzerland: Springer International Publishing.

Escrigas, Cristina, Jesús Granados Sánchez, Budd Hall, and Rajesh Tandon. 2014. "Editor's introduction. Knowledge, engagement and higher education: Contributing to social change." In *Report: Higher Education in the World*, edited by Cristina Escrigas, Jesús Granados Sánchez, Budd Hall and Rajesh Tandon. Palgrave Macmillan.

Espeland, Wendy Nelson, and Michael Sauder. 2012. "The Dynamism of Indicators." In *Governance by Indicators: Global Power through Quantification and Rankings*, edited by Kevin Davis, Angelina Fisher, Benedict Kingsbury and Sally Engle Merry, 86-109. Oxford: Oxford University Press.

Espeland, Wendy Nelson, and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York, NY: Russell Sage Foundation.

Espeland, Wendy Nelson, and Mitchell L. Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24:313-43.

Fang, Ferric C., and Arturo Casadevall. 2011. "Retracted Science and the Retraction Index." *Infection and Immunity* 79 (10):3855-9.

Fraley, R. Chris, and Simine Vazire. 2014. "The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power." *PLOS ONE* 9 (10):e109019. doi: 10.1371/journal.pone.0109019.

Heesen, Remco. 2017. "Communism and the Incentive to Share in Science." *Philosophy of Science* 84:698-716.

Hicks, Diana, and Paul Wouters. 2015. "The Leiden manifesto for research metrics." *Nature* 520:429-31.

Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* LXXXVII (1):5-22.

Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Larivière, Vincent, Véronique Kiermar, Catriona J. MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. 2016. "A simple proposal for the publication of journal citation distributions." *BioRxiv*:062109.

Larivière, Vincent, George A. Lozano, and Yves Gingras. 2013. "Are elite journals declining?" *Journal of the Association for Information Science and Technology* 65 (4):649-55.

Lee, Carole J. 2013. "The limited effectiveness of prestige as an intervention on the health of medical journal publications." *Episteme* 10 (4):387-402.

Lee, Carole J. 2015. "Commensuration bias in peer review." *Philosophy of Science* 82:1272-83.

Lee, Carole J., and David Moher. 2017. "Promote Scientific Integrity via Journal Peer Review." *Science* 357 (6348):256-7.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in peer review." *Journal of the American Society for Information Science and Technology* 64 (1):2-17.

Livneh, Ben, Eric A. Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M. Andreadis, Edwin P. Maurer, and Dennis P. Lettenmaier. 2013. "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions." *Journal of Climate* 26 (23):9384-9392.

Lozano, George A., Vincent Larivière, and Yves Gingras. 2012. "The weakening relationship between the Impact Factor and papers' citations in the digital age." *Journal of the American Society for Information Science and Technology* 63 (11):2140-45.

Macrae, C. Neil, Galen V. Bodenhausen, and Alan B. Milne. 1995. "The Dissection of Selection in Person Perception: Inhibitory Processes in Social Stereotyping." *Journal of Personality and Social Psychology* 69 (3):397-407.

Merton, Robert K. 1968. "The matthew effect in science." *Science* 1968:56-63.

Merton, Robert K. 1973. "The normative structure of science." In *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Norman W. Storer, 267-78. Chicago, IL: University of Chicago Press.

Mindner, Justin R., Philip W. Mote, and Jessica D. Lundquist. 2010. "Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains." *Journal of Geophysical Research: Atmospheres* 115. doi: https://doi.org/10.1029/2009JD013493.

National Academies of Sciences, Engineering, and Medicine,. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, D.C.: The National Academies Press.

National Science Foundation. 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. In *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*.

Nature Publishing Group. 2015. "Author Insights 2015 Survey."

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Mahlhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. "Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility." *Science* 348 (6242):1422-5. doi: 10.1126/science.aab2374.

Okasha, Samir. 2016. "On the interpretation of decision theory." *Economics & Philosophy* 32 (3):409-33.

Ormans, Laurent. 2016. "50 Journals used in FT research." accessed September 1. https://www.ft.com/content/3405a512-5cbb-11e1-8f1f-00144feabdc0.

Ridgeway, Cecilia L., and Shelley J. Correll. 2006. "Consensus and the creation and status beliefs." *Social Forces* 85 (1):431-53.

Rubin, Hannah, and Cailin O'Connor. 2018. "Discrimination and Collaboration in Science." *Philosophy of Science* 85:380-402.

San Francisco Declaration on Research Assessment. 2013. "The San Francisco Declaration on Research Assessment (DORA)." accessed September 1. https://sfdora.org/read/.

Sauder, Michael, and Wendy Nelson Espeland. 2006. "Strength in numbers? The advantages of multiple rankings." *Indiana Law Journal* 81 (1):205-27.

Schimmack, Ulrich. 2015. "Replicability Ranking of 26 Psychology Journals." January 18. https://replicationindex.wordpress.com/2015/08/13/replicability-ranking-of-26-psychology-journals/.

Science. "Mission and Scope." accessed September 1. http://sciencemag.org/about/mission-and-scope.

Strevens, Michael. 2003. "The role of the priority rule in science." *Journal of Philosophy* 100 (2):55-79.

West, Jevin D., Theodore C. Bergstrom, and Carl T. Bergstrom. 2010. "The Eigenfactor Metrics™: A network approach to assessing scholarly journals." *College & Research Libraries* 71 (3):236-44.

Willer, Robb, Ko Kuwabara, and Michael W. Macy. 2009. "The False Enforcement of Unpopular Norms." *American Journal of Sociology* 115 (2):451-90.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. 2015. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.

Wilsdon, James, Judit Bar-Ilan, Robert Frodeman, Elisabeth Lex, Isabella Peters, and Paul Wouters. 2017. Next-generation metrics: Responsible metrics and evaluation for open science. Report of the European Commission Expert Group on Altmetrics. European Commission.

Zollman, Kevin J. S. 2018. "The Credit Economy and the Economic Rationality of Science." *The Journal of Philosophy* 115:5-33.

Zuckerman, Harriet, and Robert K. Merton. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System." *Minerva* 9 (1):66-100.

Pragmatism and the content of quantum mechanics

Peter J. Lewis

Draft – please don't quote

Abstract

Pragmatism about quantum mechanics provides an attractive approach to the question of what quantum mechanics says. However, the conclusions reached by pragmatists concerning the content of quantum mechanics cannot be squared with the way that physicists use quantum mechanics to describe physical systems. In particular, attention to actual use results in ascribing content to claims about physical systems over a much wider range of contexts than countenanced by recent pragmatists. The resulting account of the content of quantum mechanics is much closer to quantum logic, and threatens the pragmatist conclusion that quantum mechanics requires no supplementation.

1. Introduction

Quantum mechanics is, notoriously, a theory in need of interpretation. But there is very little agreement on what kind of interpretation it needs. That is, there is very little agreement concerning what the foundational problems of quantum mechanics *are*, and without such agreement, there is little hope for a consensus concerning what an acceptable solution to the problems might look like.

Here is a way to divide up the territory. We can distinguish between *descriptive* and *normative* questions concerning quantum mechanics. Descriptive questions concern what quantum mechanics *says*—the *content* of the theory, as expressed in textbooks and used in labs. Normative questions concern what quantum mechanics *should* say—and in particular, whether it should say something different from what it actually does say.

All parties to the debates over the foundations of quantum mechanics would agree, I think, that there is a legitimate descriptive question concerning the content of quantum mechanics. Even those philosophers and physicists who think that quantum mechanics wears its interpretation on its sleeve at least feel the need to correct the mistaken impressions of *other* philosophers and physicists concerning what quantum mechanics says. The normative question presupposes an answer to the descriptive one: some think quantum mechanics is just fine the way it is, others contend that it needs to be replaced or supplemented with something radically different, and in large part this difference in attitude depends on prior differences concerning the answer to the descriptive question.

As an illustration, consider a fairly standard narrative concerning the descriptive and normative questions. Descriptively speaking, quantum mechanics depends on a distinction between measurements and non-measurements: measurements follow one dynamical law, the collapse dynamics, and non-measurements follow a different dynamical law, the Schrödinger dynamics. Since these two dynamical processes are incompatible, a precise formulation of quantum mechanics requires a precise dividing line between measurements and non-measurements. Quantum mechanics nowhere provides such a thing—and indeed, it seems highly unlikely that a term like “measurement” could be given a physically precise definition. So

descriptively speaking, quantum mechanics is inadequate as a physical theory. On the basis of this measurement problem, Bell (2004, 213–231) recommends replacing quantum mechanics with either a pilot-wave theory or a spontaneous collapse theory. For similar reasons, Wallace (2012, 35) recommends replacing quantum mechanics with a many-worlds theory.[1]

But not everybody concurs. There are alternative narratives according to which quantum mechanics, descriptively speaking, is just fine as it is, and hence there is no normative pressure to supplement or replace it. One prominent version proceeds from the quantum logic of von Neumann (1936) and Putnam (1975) through to the quantum information theory of Bub (2016). According to this approach, quantum mechanics describes a non-classical event space—in terms of truth values, a non-Boolean algebra, and in terms of probability ascriptions, a non-simplex distribution. No-go theorems (arguably) show that it is impossible to construct a set of events obeying classical Boolean logic or classical Kolmogorov probability that reproduces the empirical predictions of quantum mechanics. The implication is that in quantum mechanics we have discovered something important about the fundamental event structure of the world. Seeking to replace or supplement quantum mechanics with a theory obeying classical logic and classical probability theory amounts to a quixotic attempt to impose a structure on the world that it manifestly does not have (Bub 2016, 222). The measurement problem, on this account, results from a mistaken demand for a dynamical explanation of the individual events in the quantum structure, when no such explanation is available (Bub 2016, 223)

[1] Wallace takes the many-worlds theory to be a precise statement of the content of quantum mechanics, rather than a replacement for it. I take up the question of whether the many-worlds structure is present in quantum mechanics as it stands in section 2.

This fundamental difference of opinion—between those who take the measurement problem seriously and those who regard it as a pseudo-problem—continues to divide the foundations of physics community today. Hence the descriptive question—the question of what quantum mechanics actually *says*—remains a pressing one. In this paper, I argue for a particular way of approaching the descriptive question. The methodology is the pragmatist one of Healey (2012; 2017) and Friederich (2015), but the answer to the descriptive question that results from following this methodology, I argue, differs in an important way from the answers that Healey and Friederich give. I conclude by assessing the consequences of this answer to the descriptive question for the normative question.

2. The descriptive question

So how should we approach the descriptive question? Consider a straightforward realist approach to the content of scientific theories. A theory, at least in physics, is typically expressed using a particular mathematical structure. The *state* of a physical system is generally identified with a mathematical entity that resides in a particular abstract space, and the *dynamics* of the theory tell us how that state evolves over time. So, for example, in many applications of classical mechanics, the state of a physical system can be represented by a set of vectors in a three-dimensional Euclidean space, and the dynamical laws of Newtonian mechanics tell us how the set of vectors evolves over time. The interpretation of the mathematics is fairly straightforward: the vectors represent the positions and momenta of point-like particles, and classical mechanics tells us how the properties of the particles change.

Such an approach can equally be applied to quantum mechanics (Albert 1996). According to quantum mechanics, the state of a physical system is identified with a complex-valued function defined on a configuration space—a space with three dimensions for each particle in the system. A dynamical law, the Schrödinger equation, tells us how this function, the wave-function, changes over time. Then by analogy with classical mechanics, the wave-function must be a representation of the physical properties of the quantum system as they change over time.

The continuity with classical mechanics in the above account is attractive, but there are surprising consequences. For an *N*-particle system, the wave-function is defined over a 3*N*-dimensional configuration space, and it cannot be represented without loss in a three-dimensional space. This has led some to conclude that a straightforward realist reading of quantum mechanics shows that the three-dimensionality of our physical world is illusory (Albert 1996). Furthermore, if we model a measurement using quantum mechanics, the wave-function ends up with components corresponding to each possible outcome of the measurement—not just one outcome, as is the case classically. This leads Everettians like Wallace (2012) to conclude that a straightforward realist reading of quantum mechanics shows that every possible outcome of a measurement actually occurs.

These conclusions might be right, but do they simply follow from close attention to the structure of quantum mechanics? There are reasons to be suspicious. As Healey (2017, 116) notes, conclusions of this kind depend on the assumption that the wave-function plays the same descriptive role in quantum mechanics as the position-momentum vectors play in classical mechanics. If this assumption is itself up for grabs in the interpretation of quantum

mechanics, then neither of these conclusions is warranted. But how do we adjudicate the question of whether the wave-function describes physical systems or whether it has some other, non-descriptive role? Is there a metaphysically neutral methodology that could be used to answer this question? Healey (2012; 2017) and Friederich (2015) think that there is.

3. Pragmatism

Consider an analogy. "Stealing is bad" has the same grammatical structure as "Cherries are red". But it is far from clear that both sentences should be taken as descriptive. In particular, badness, taken as a property of actions, seems like a queer kind of property, imperceptible and disconnected from the other properties of the action. Expressivists seek to dissolve the problem of the nature of badness by claiming that a sentence like "Stealing is bad" should be taken as expressive rather than descriptive—as expressing our attitude towards stealing. Pragmatists further coopt expressivism as a variety of pragmatism (Price 2011, 9). Pragmatists stress the variety of uses of language, noting that sentences with superficially similar form can be used in radically different ways. "Cherries are red" is used to describe a class of objects, whereas "Stealing is bad" is used to express our attitude towards a class of actions.

Pragmatism, then, enjoins us to pay close attention to how a sentence is *used* in order to find out what it means. Healey (2012; 2017) and Friederich (2015) each suggest that the pragmatist approach provides us with a metaphysically neutral methodology for probing the content of quantum mechanics. That is, we can look at how various quantum mechanical claims are used by physicists in order to determine what those claims mean. This strikes me as a welcome suggestion. In the rest of this section I present the conclusions of their pragmatist

inquiries; in the next, I consider whether the language use of physicists actually supports those conclusions.

Healey (2012) distinguishes between *quantum claims* and *non-quantum magnitude claims*. The former explicitly mention quantum states, quantum probabilities, or other novel elements of the theory of quantum mechanics. The latter are claims about the magnitude of a physical quantity that do *not* involve quantum states, quantum probabilities etc. In keeping with the pragmatist methodology, Healey bases this distinction on the way the two kinds of claims are used. Non-quantum magnitude claims are used in a straightforwardly descriptive way. But quantum claims are used in a different way: they are used, not to *describe* a system, but to *prescribe* a user's degrees of belief in various non-quantum magnitude claims.

As an example, Healey appeals to the Interference experiments of Juffmann et al. (2009), in which C_{60} molecules are passed through an array of slits and then deposited on a silicon surface. To derive quantum mechanical predictions for this experimental arrangement, quantum states are ascribed to C_{60} molecules. That is, quantum claims of the form "The molecule has state $|\psi\rangle$" are used, via the Born rule, to ascribe probabilities to claims concerning the various possible locations of the molecules on the silicon surface. These latter claims—of the form "The molecule is located in region R"—are non-quantum magnitude claims. The job of the non-quantum magnitude claims is to describe the physical system, but the job of the quantum claims is to prescribe degrees of belief in the non-quantum magnitude claims for an appropriately situated observer. In this respect Healey's approach is like the expressivist's in ethics: claims that have superficially similar grammatical forms have very different functions.

Another important strand in the pragmatist approach concerns the role of decoherence. After the C_{60} molecule hits the silicon surface, complicated interactions with the surface mean that the state of the molecule-environment system becomes approximately diagonal when written as a density matrix in the position basis. This in turn insures that the probabilities ascribed by the Born rule to various claims about the molecule's position closely obey the probability axioms. But before the molecule encounters the silicon surface, its state is a coherent superposition—a state that is not even approximately diagonal, and for which the Born rule does not ascribe probabilities to location claims that closely obey the probability axioms. For such a state, the Born rule does not prescribe appropriate degrees of belief in the non-quantum location claims, and so assertion of such claims prior to decoherence is not *licensed* by quantum mechanics. Decoherence, then provides a demarcation between situations in which it is appropriate to have a well-defined degree of belief in a non-quantum magnitude claim, and situations in which it is not.

The central finding of the Healey-Friederich pragmatist approach is that attention to the use of quantum mechanical language shows that claims about the quantum state of a system are not used to describe that system. Hence, we should not think of the wave-function as a representation of the physical properties of the quantum system as they change over time. This perspective has the advantage that the measurement problem does not arise: if the wave-function doesn't represent the system, then we don't have to worry that the dynamical laws for wave-function evolution are different for measurements and non-measurements. In fact, if the quantum state is prescriptive, then the difference between measurements and non-

measurements arises quite naturally: the results of measurements have a direct and obvious influence on what you should believe.

Hence the pragmatist approach provides a clear answer to the descriptive question: quantum mechanics, in itself, says *nothing* about the world. As Healey (2017, 12) puts it, "quantum theory has no physical ontology". Rather, quantum mechanics tells us what to believe about non-quantum ontology—about particles, or in the case of quantum field theory, about fields. Furthermore, this answer to the descriptive question suggests an answer to the normative question: since the measurement problem doesn't arise, there is no motivation for supplementing or replacing quantum mechanics with something else.

4. Actual use, counterfactual content

Thus far, I have said little about the evidence that backs up Healey's claims about how quantum claims and non-quantum magnitude claims are used. Indeed, direct evidence from the language use of physicists is likely to be unenlightening: that a claim is asserted in a given context provides no direct evidence concerning whether its content is descriptive or prescriptive.

To fill this gap, Healey appeals to an inferentialist account of the link between use and meaning derived from the work of Robert Brandom (2000): the meaning of a claim is identified with the set of material inferences it licenses. So by looking at the way a claim is used in licensing inferences, we can gain evidence about what it means. And here the distinction between prescriptive quantum claims and descriptive non-quantum magnitude claims seems to be well motivated. In the practice of physics, a claim about the quantum state of a system is

used to infer Born probabilities, and nothing more. If Born probabilities are taken to be rational degrees of belief, then the prescriptive content of a quantum claim exhausts its meaning.

A non-quantum magnitude claim, on the other hand, can license a wide variety of inferences. From the claim that a C_{60} molecule is located in a particular region of the silicon surface, we can infer that an electron microscope will produce an image of the molecule if directed at that region (Juffmann et al. 2009, 2). We can infer that if the silicon surface is left untouched for two weeks, the C_{60} molecule will remain in the same place (Juffmann et al. 2009, 2). Under suitable conditions, we can infer that the C_{60} molecule will emits photons; under different conditions, that it will act as a nucleation core for molecular growth (Juffmann et al. 2009, 3). In other words, the inferences licensed by the non-quantum magnitude claim support the interpretation that the meaning of the claim is descriptive rather than merely prescriptive.[2]

So there is a good case to be made, I think, that actual use supports the distinction between prescriptive quantum claims and descriptive non-quantum magnitude claims. But there is a further strand to the Healey-Friederich interpretation, namely that non-quantum magnitude claims are only licensed after decoherence. This claim, I think, does not stand up so well to scrutiny.

Consider C_{60} interference again. After the molecule has adhered to the silicon surface, the state of the molecule is decoherent, and the claim that the molecule has a particular

[2] There is a sense in which the meaning of *any* claim is prescriptive according to the inferentialist program: the claim about the location of the molecule licenses an inference to a certain *degree of belief* that the electron microscope will produce an image of it. But still, there is a reasonable distinction here: the quantum claim licenses inferences only via the Born rule, whereas the non-quantum magnitude claim licenses inferences via a huge variety of schema typical of small physical objects. The latter is just what it is for a claim to be descriptive.

location is licensed—that is, it is appropriate to associate a particular degree of belief with the claim, and if that degree of belief is high enough, it is appropriate to assert the claim. But before the molecule has adhered to the silicon surface, the state of the molecule is coherent, and no claim about the location of the molecule is licensed—it is not appropriate to associate a degree of belief with such a claim, or to assert it. Similar considerations apply to properties other than location.

This seems to fly in the face of actual use. For example, in the description of the C_{60} interference experiment, Juffmann et al. (2009, 2) assert that "all transmitted particles arrive with the same speed," and "about 110cm behind the source, the molecules encounter the first diffraction grating," apparently ascribing both speed and location to C_{60} molecules prior to decoherence. This doesn't seem to be an isolated incident: physicists routinely talk of preparing, selecting, spraying, shooting and trapping particles, ions and molecules, and this talk typically involves making claims about these objects prior to any eventual decoherence.

It is possible, of course, that this is just "loose talk", or an indirect way of making claims about the quantum state of the systems concerned. But given the frequency of such claims, and given the reliance of the pragmatist methodology on *use*, this seems like a shaky game to play. It would be better, all things considered, if such claims could be accommodated within the pragmatist interpretation, rather than explained away as anomalies.

But there are obvious barriers to licensing non-quantum magnitude claims prior to decoherence. As Friederich (2015, 79) notes, the Born rule is only "reliable" when applied to decoherent states, in the sense that only for such states are the numbers it produces guaranteed to closely obey the probability axioms. Given some reasonable assumptions about

rationality, it is plausible that numbers that do not closely obey the probability axioms could not be rational degrees of belief. Furthermore, Healey argues that asserting a non-quantum magnitude claim prior to decoherence is likely to be misleading. For example, suppose one asserts (with Juffmann et al.) that "about 110cm behind the source, the molecules encounter the first diffraction grating." One might infer from this that each molecule passes through exactly one slit in the grating, and hence that the presence of the other slits is irrelevant, and hence that there is no possibility of interference (Healey 2012, 745).

So the pragmatist approach seems to face a dilemma: either it fails to accommodate the actual language use of physicists, or it licenses misleading assertions and irrational degrees of belief. Isn't there another way? I think there is. Consider a mundane claim like "There is beer in the fridge." In typical contexts, an assertion of this claim licenses the inference that if you were to go to the fridge and open the door, you could take a beer and drink it. Of course, you might not actually do this; maybe you don't want a beer. That is, the inference here is a counterfactual one. A good deal of the inferential content of our assertions has this counterfactual character.

Now return to the quantum context. Consider again the claim that "about 110cm behind the source, the molecules encounter the first diffraction grating." What content could that claim have? If we broaden the notion of inferential content to include counterfactual inferences, then the content seems fairly clear: if we were to replace the first diffraction grating with a detector taking up the same region of space, then the Born rule would ascribe a degree of belief close to 1 to detecting the molecules.

How does the inclusion of counterfactual content avoid the barriers to licensing non-quantum magnitude claims prior to decoherence? Note that the counterfactual content of the claim about the molecules involves a counterfactual intervention on the system—a counterfactual measurement. The counterfactual measurement induces counterfactual decoherence. The Born probabilities are conditional on this intervention and the associated decoherence, so the Born probabilities for various position claims concerning the molecules are not, after all, unreliable, in the sense of violating the probability axioms.

Neither should there be any danger of being misled by an assertion that the C_{60} molecules encounter the grating, because the counterfactual conditions implicit in the content of that assertion are distinct from the conditions that actually obtain in the apparatus. That you *could* detect the molecules at the diffraction grating, given a different experimental arrangement, doesn't license the inference that there *is* no interference, given the actual experimental arrangement. Admittedly, though, this amounts to a weakening of the content of position claims from the classical case, as spelled out in the next section.

5. A happy convergence?

I have argued that non-quantum magnitude claims have assertible content in a far wider range of contexts than countenanced by Healey or Friederich. If there is some counterfactual intervention on a system that would produce decoherence in the basis defined by a given observable, then claims about the values of that observable have content. And since counterfactual interventions only have to be realizable in principle, this means that claims about the value of an observable for a system *generally* have content, whether or not the

system *actually* decoheres in the basis defined by that observable. This has the welcome consequence that the frequent assertions made by physicists about the properties of systems prior to decoherence are contentful.

A potential cost of such permissiveness about content is that the structure of this content is, in general, non-Boolean. Consider again a C_{60} molecule that is approaching the first diffraction grating, and consider an assertion of "The molecule passes through the leftmost slit". This assertion has content, on the proposed view, because in principle there is an intervention on the system that would produce decoherence in a basis defined by an observable that distinguishes which slit the molecule passes through. Still, assertion of the claim would not be appropriate, simply because there are many slits in the grating, so the Born rule ascribes it a low probability. The same goes for every other slit in the grating. Nevertheless, the assertion that "The molecule passes through the leftmost slit, or the second to the left, or…" is assertible, since the Born rule ascribes it a probability close to 1. The disjunction is assertible, but none of the disjuncts is assertible. Since assertibility is a surrogate for truth in the pragmatist context, this is equivalent to saying that the disjunction is true, but none of the disjuncts is true.

One might take this to be unacceptable on the pragmatist view—especially if you endorse an inferentialist pragmatism, as Healey does. From a disjunctive claim you can straightforwardly infer that at least one of the disjuncts is true. If the content of a claim is identified with the inferences that it licenses, then part of the meaning of the disjunctive claim about the C_{60} molecule is that some assertion of the form "The molecule went through slit *x*" is true. Hence my proposal about content threatens to violate the inferentialist account of

meaning. The pragmatist interpretation of Healey and Friederich avoids this problem by insisting that claims about systems have meaning only after suitable decoherence.

Of course, pragmatism is not necessarily tied to an inferentialist account of meaning. But even given inferentialism, there is arguably no real problem here. Physicists are *selective* in the inferences they draw: from the disjunctive claim, they don't infer that the C_{60} molecule goes through some particular slit, so they don't infer a lack of interference. But they do infer that the molecule will arrive at the silicon surface, that it might radiate a photon in flight, and so forth. That is, the inferences drawn by physicists from their claims about pre-decoherent systems suggest that the non-Boolean structure of those claims is already *built in* to the meanings associated with those claims and revealed in inference.

This suggests that close attention to the way non-quantum magnitude claims are actually used leads to a happy convergence between pragmatism and the quantum logical approach. Physicists assert claims about particles even when the state does not decohere, and such claims seem to be meaningful. But physicists are not inclined on that basis to draw all the inferences that a full Boolean structure to their claims would license. Quantum mechanics apparently weakens the meaning of many claims about pre-decoherent physical systems, but without rendering those claims meaningless.

6. The normative question

As a methodology for addressing the *descriptive* question of the content of quantum mechanics, the pragmatist approach seems entirely appropriate: look to the *use* of physicists to determine what the various claims involved in the theory mean. At the hands of Healey and

Friederich, this approach yields the important insight that while non-quantum magnitude claims are used to describe physical system, quantum claims are used to prescribe appropriate degrees of belief in non-quantum magnitude claims. But Healey and Friederich go further, in limiting the assertibility of non-quantum magnitude claims to contexts in which the quantum state is decoherent in the relevant basis. This, I have argued, cannot be squared with the actual use of such claims. I propose instead that non-quantum magnitude claims *generally* have well-defined content, understood in terms of a counterfactual intervention on the system. This change to the pragmatist approach means that it ends up looking a lot like the quantum logical approach that preceded it. Indeed, the pragmatist approach might be regarded as a *justification* for quantum logical claims concerning the content of quantum mechanics.

But where does all this leave the *normative* question concerning whether quantum mechanics is fine as it is, or whether it should be supplemented or replaced? Healey and Friederich argue that quantum mechanics is fine as it is: if quantum claims do not describe physical systems, then there can be no conflict between the way that quantum mechanics describes systems during measurements and the way it describes them during non-measurements. If there is no measurement problem, then there is no motivation to replace such a successful theory. If, as Healey (2017, 12) maintains, quantum theory "states no facts about physical objects or events," then there can be no requirement that we come up with an *explanation* of quantum facts and events.

However, I have suggested that quantum theory has more content than the pragmatists countenance. In one sense, I agree that quantum theory states no facts: a quantum claim, such as the attribution of a quantum state to a system, is not a description. But in another sense,

there are distinctive quantum facts, or at least facts with a distinctive quantum structure: non-quantum magnitude claims about pre-decoherent systems exhibit the non-Boolean structure characteristic of quantum mechanics. This is the sense in which quantum logic gets things right.

Notably, though, the proponents of quantum logic *also* often take the view that quantum logic dissolves the measurement problem (e.g. Putnam 1975, 186). But this dissolution is widely regarded to be a failure (e.g. Bacciagaluppi 2009, 65) Once one has admitted that the structure of true (i.e. assertible) claims for a quantum system is non-Boolean, the question of *how* the world manages to instantiate this structure becomes legitimate and pressing. A denial that any explanation is required looks suspiciously like instrumentalism. And since any answer to this question goes beyond quantum mechanics as it stands, the call for explanation involves a demand to supplement quantum mechanics, or to replace it with something more fundamental.

Of course, given the no-go theorems, the path to an explanation of the structure of quantum facts is by no means clear. But neither do the no-go theorems show that an explanation is *impossible* (Friederich 2015, 161).[3] If the foregoing is correct, then pragmatism is an excellent way to *expose* the foundational problems of quantum mechanics, but it is not a means to *dissolve* them.

References

[3] Interestingly, Friederich (2015, 161) suggests supplementing quantum mechanics with sharp values for all observables, even though this seems at odds with his therapeutic aim of dissolving the foundational problems of quantum mechanics rather than solving them (2015, 6).

Albert, David Z. (1996), "Elementary quantum metaphysics," in J. T. Cushing, A. Fine and S. Goldstein (eds.), *Bohmian Mechanics and Quantum Theory: An Appraisal*. Dordrecht: Springer, 277-284.

Bacciagaluppi, Guido (2009), "Is logic empirical?" in K. Engesser, D. M. Gabbay and D. Lehmann (eds.), *Handbook of Quantum Logic and Quantum Structures.* Amsterdam: North-Holland, 49-78.

Bell, J. S. (2004), *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press.

Brandom, R. (2000), *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Bub, Jeffrey (2016), *Bananaworld: Quantum Mechanics for Primates*. Oxford: Oxford University Press.

Friederich, Simon (2015), *Interpreting Quantum Theory: A Therapeutic Approach*. Basingstoke: Palgrave Macmillan.

Healey, Richard (2012), "Quantum theory: a pragmatist approach," *British Journal for the Philosophy of Science* 63: 729-771.

Healey, Richard (2017), *The Quantum Revolution in Philosophy*. Oxford: Oxford University Press.

Juffmann, T., Truppe, S., Geyer, P., Major, A. G., Deachapunya, S., Ulbricht, H., and Arndt, M. (2009), "Wave and particle in molecular interference lithography," *Physical Review Letters* 103: 263601.

Price, Huw (2011), *Naturalism Without Mirrors*. Oxford: Oxford University Press.

Putnam, Hilary (1975), "The logic of quantum mechanics," in *Mathematics, Matter and Method: Philosophical Papers Volume 1*. Cambridge: Cambridge University Press.

von Neumann, John (1932), *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer-Verlag.

Wallace, David (2012), *The Emergent Multiverse*. Oxford: Oxford University Press.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -436-

Tool Migration: A Framework for Analyzing Cross-disciplinary Use of Mathematical Constructs

Chia-Hua Lin
University of South Carolina / KLI

Abstract. Mathematical formalisms that are constructed for inquiry in one disciplinary context are sometimes applied to another, a phenomenon that I call 'tool migration.' Philosophers of science have addressed the advantages of using migrated tools. In this paper, I argue that tool migration can be epistemically risky. I then develop an analytic framework for better understanding the risks that are implicit in tool migration. My approach shows that viewing mathematical constructs as tools while also acknowledging their representational features allows for a balanced understanding of knowledge production that are aided by the research tools migrated across disciplinary boundaries.

Keywords: Cross-disciplinarity, tool migration, epistemic risks

1. Introduction

Mathematical formalisms that are constructed for scientific inquiry in one disciplinary (or sub-disciplinary) context are applied to another. Philosophers of science have started paying attention to this cross-disciplinary aspect of scientific practice. For instance, the discussion of 'model transfer' concerns a relatively small set of mathematical models that are applied in multiple disciplinary contexts. Humphreys (2004) proposes that models that are transferred to study phenomena of a different domain owe their versatility to the computational tractability they afford. In contrast, Knuuttila and Loettger (2014, 2016) suggest that in addition to tractability, versatile models also offer conceptual frameworks for theorization, which they label 'model templates.' However, these analyses do not deal with the risks inherent in this aspect of scientific practice. Consider the use and development of game theory in evolutionary biology as an example. In importing game theory, which was originally conceived to describe strategic interaction between rational agents typically studied by social scientists, evolutionary biologists may need to modify the theory in order to generate knowledge about presumably non-rational agents, at least in many cases. One can then assume that any changes to the theory--between its established applications in social sciences and its novel uses in evolutionary biology--require special attention so as to avoid misinterpreting an analysis.

Despite the advantages, there might be risks associated with using mathematical constructs across disciplines. In this paper, I ask: might there be patterns of transfer that may undermine the effectiveness of the imported mathematical formulation? What would these

patterns, if any, look like? This paper is an attempt to explore the conditions in which importing mathematical constructs may be epistemically risky. To begin, I develop a framework to systematically characterize the landscape of mathematical importations. The goal of such a framework is two-fold. Proximately, the framework captures characteristics of migration that the current terminology, such as 'model transfer' or 'importing/exporting,' fails to discern. Ultimately, with this additional discernibility, I suggest that one may start to explore and identify patterns of importation that may be subject to epistemic risks, such as misinterpretation of an outcome produced by using a imported mathematical construct.

In Section 2, I argue that one can view mathematical constructs in science in terms of 'research tools' and that transporting such tools across disciplines, which I call 'tool migration,' can in some cases be a disservice to science. Next, I classify tool migration based on two kinds of contextual details that bear significance to the effectiveness of the migrated research tool in a foreign context. In Section 3, I apply this approach to the use and development of game theory in evolutionary biology. Finally, in Section 4, I discuss in what ways this tool migration framework, which is essentially a typology of four types of tool migration, may help to characterize epistemically risky patterns of tool migration.

2. Theoretical Background

Although the notion of epistemic risks associated with migration of mathematical constructs has not been explicitly addressed, the idea of viewing mathematical constructs as research tools follows from the discussion on the ontology of scientific models. Ever since the shift of attention to scientific practice (e.g., Hacking 1983), there has been a growing literature in which models in science are viewed as entities *detachable* from theory and data (e.g., Morrison 1999; Morgan and Morrison 1999). One recent predecessor to my tool migration account is a pragmatic approach to scientific models put forth by Boon and Knuuttila (2008). In their paper, which uses examples from engineering, they argue that scientific models are better understood as 'epistemic tools' instead of as representations of some target systems in the world. Boon and Knuuttila's argument draws heavily on the epistemological roles of scientific models in relation to the scientists who use them. According to them, scientific models allow their users "to understand, predict, or optimize the behavior of devices or the properties of diverse materials" (2008, 687). Thus, for an ontological account of scientific models to be productive and realistic, as they argue, it should be sensitive to the relation between the models and the modelers, i.e., the tools and their users. An adequate evaluation of Boon and Knuuttila's argument will take us far afield, but my work will show that both the representational and the pragmatic aspects are indispensable to a better understanding of the epistemic risks in tool migration.

2.1 Viewing mathematical constructs as research tools

In general terms, any mathematical construct that is to be *used or operated* in an algorithmic manner, and the outcome of whose operation is to be *interpreted* in order to answer a research question, is an example of what I am calling a research tool. Let me first unpack the operational aspect of a research tool.

Let's assume that the proper use of any mathematical constructs employed in scientific research is expected to produce consistent results. To achieve this consistency, then, a well-defined procedure needs to accompany such a construct so that anyone who follows the procedure expects, and is expected, to obtain the same outcome given the same input. For instance, when performing a game-theoretic analysis, one goes through a sequence of steps, such as: (i) identify the players and the acts available to them, (ii) identify the payouts in every set of acts, (iii) find the 'Nash equilibria,' which refers to a set of acts, one for each player, in which no player could improve his or her payoff by unilaterally changing act. A similar algorithmic procedure can be seen when applying, say, Newton's law of gravitation:

$$F_{grav} = G\frac{m_1 m_2}{r^2}. \qquad (1.1)$$

For example, the sequence of steps to obtain the magnitude of the gravitational force, F_{grav}, between any two objects includes: (i) identify the mass of each object, (ii) identify the distance between them, (iii) complete the equation in which 'm_1' and 'm_2' refer to the masses of the two objects, 'r' the distance in between, and 'G' the gravitational constant. In these two examples, when the first two steps produce consistent input, the third step is expected to generate the same output.

Moreover, concerning the interpretational aspect of a research tool, the output of a series of symbol assignments and manipulations can be understood *only through the lens of some interpretation*. The Nash-equilibrium of a game is a meaningful 'solution' in virtue of the usual understanding of the game-theoretic formulation of a problem. Similarly, the meaning of the value obtained through completing the equation in (1.1) is derived from the usual interpretation of the quantities appearing in the equation and the theoretical context in which those quantities are defined.

Finally, assume that something can be viewed as a tool if it serves as a means to an end. In this case, then, mathematical constructs like game theory or mathematical formulas can be seen as research tools. In the case of applying a mathematical construct, the goal of performing a sequence of prescribed steps goes beyond merely completing the calculation and obtaining a result. Instead, the output is to be interpreted so that one may solve a problem, answer a research question, or gain knowledge about a subject-matter. Thus, a mathematical construct that prescribes algorithmic symbol manipulation can be seen as a research tool, assisting its users to meet an end. Manipulating symbols is a means to the end that was specified during the mathematical formulation of the research problem.

2.2 Epistemic risks of tool migration

Another predecessor to my account is Morgan's discussion of the re-situating of knowledge (2014). According to her, knowledge production is necessarily 'situated,' and consequently, applying a piece of knowledge outside its initial context requires effort - different contextual situations require different 're-situating' strategies. The term 're-situation' thus captures what scientists do in practice to transport locally generated knowledge across contexts. As she argues, to make an instance of scientific knowledge accessible outside its production site, one needs to establish inferential links between the production site and the destination site. However, she suggests, whether a re-situation of knowledge contributes to scientific progress depends on whether the transport secures some sort of inferential safety.

Building from Morgan's notion of the re-situation of knowledge, I argue that cross-disciplinary use of research tools is epistemically risky. Given the locality of scientific knowledge production, applying scientific knowledge outside its production site may come with epistemic risks. For example, between the production site and a destination site, there may be incongruent disciplinary characteristics (e.g., implicit theoretical assumptions) that fail to be captured by the inferential strategy, such that knowledge from the former cannot be transferred to the latter. Similarly, we can assume that the construction of a research tool is also *situated* in nature. Namely, a research tool is conceived to be operated and to extend our knowledge concerning a subject-matter *given a particular disciplinary context*. It follows that cross-disciplinary use of research tools is as epistemically risky as re-situating knowledge. That is, the epistemic reliability (i.e., general ability or tendency to produce knowledge) of some research tool in one disciplinary context does not necessarily carry over to another.

The concept of 'tool migration' captures both the 'situated-ness' of a research tool that was established in its native discipline and the effort it takes to 're-situate' the tool in a foreign discipline. Naturally, in the process of uprooting a research tool, significant contextual details—ranging from implicit expertise to important background assumptions—may be stripped away. Likewise, during re-situation, new features may be introduced to the tool so as to treat a different subject matter in a new disciplinary context. Together, due to the possibility of losing or gaining significant contextual details, or both, a cross-disciplinary tool migration risks undermining the effectiveness of the tool. These risks include, for example, misinterpretation of the research result or failure to produce genuine knowledge. Thus, it follows that tool migration can in some cases be a disservice to the production of knowledge.

Acknowledging these challenges, some have argued against the cross-disciplinary effort to integrate disciplinary knowledge (e.g., van der Steen 1993). Alternatively, one might try to overcome these challenges so long as the risks are better understood and managed. To understand the risks, I suggest that we first look at the patterns of tool migration. Among these patterns, we might find that some of them could be epistemically risky. Having established the

notions of research tools and risks involved with tool migration, I turn to the contextual details that are closely related to a tool's epistemic performance.

2.3 Contextual details of a research tool: the target profile and the usage profile

The construction of a research tool is necessarily situated within a context. In order to compare and contrast between the native (or established) context and the foreign context of a migrated tool, I single out two major types of details.

The first type concerns the assumptions about the entities that are studied by a subject-matter for which the tool is developed. For instance, game theory defines what it considers as a game, a player, or an act. For simplicity, I call *all* the assumptions that a tool makes about its target entities the tool's 'target profile.'

The second type considers *the ways* in which one interprets the output from applying a tool in his or her research. In a game-theoretic analysis, for example, by following an algorithmic procedure, one obtains a solution of a game in the form of a Nash equilibrium. Depending on the game that one was analyzing, the solution could be understood as an explanation of economic behavior, or a prediction about it, or it could be used to optimize an strategic interaction. For simplicity, I call *all* the ways in which a tool is intended to be used, e.g., describing, predicting, optimizing, or explaining its target phenomenon, the tool's 'usage profile.'

Together, as I demonstrate in Section 4, the 'target profile' and 'usage profile' allow one to detect patterns of changes in the contextual details between the established use and the novel use of a research tool. They are able to do this because these two profiles offer a coarse resolution; looking through the lens of the target profile and usage profile, one zooms out from particular cases of tool migration so as to detect patterns of cross-disciplinary transport. Further analyses of these patterns will then shed lights on their associated epistemic risks.

2.4 Four types of tool migration

With the two profiles of a research tool and the two contexts in which the tool is used, i.e., a novel use and an established use, one can distinguish four types of tool migration.

First, compared to its established use, when a novel use of a tool catalyzes changes in both target and usage profiles, the tool migration is transformative, and therefore I call it a ***tool-transformation***. Second, in contrast, when both target and usage profiles remain more or less intact after the migration, the tool's novel use is considerably similar to its previous applications. Thus, I call such a case ***tool-application***. Between these two extreme types, there are novel uses of a research tool that alter only one of the two profiles but not both. When a tool changes its target profile but not its usage profile, I call it a ***tool-transfer,*** and when a tool changes its usage profile but not the target profile, I call it a ***tool-adaptation***. See **Table 1** for a summary.

Table 1
A Typology of Tool Migration

Between established and novel uses of a research tool	**Usage profile remains**	**Usage profile deviates**
Target profile remains	'Tool-application'	'Tool-adaptation'
Target profile deviates	'Tool-transfer'	'Tool-transformation'

Among these four types of tool migration, tool-transfer is arguably the most familiar to the philosophers of science. Humphreys coins the term 'computational templates' to refer to a relatively small number of mathematical equations that are applied to investigate different domains of phenomena (2002, 2004). Bailer-Jones (2009) discusses such a scientific practice in terms of mathematical analogy. For one example, Newton's law of gravitation was intentionally sought after to model electrostatic force (see Bailer-Jones 2009 for a detailed account). The important parallel between the two formulas, shown in (1.2), is that both types of forces (gravitational and the electrostatic) are proportional to the inverse of the square of the distance, r, between two masses, m_1 and m_2, or two charges, q_1 and q_2. The constants that appear in both formulas scale the quantities to match empirical phenomena.

$$F_{grav} = G\frac{m_1 m_2}{r^2} \quad \text{and} \quad F_{el} = k\frac{q_1 q_2}{r^2} \tag{1.2}$$

In contrast, the other three types of tool migration, despite prominent examples, are less explored in regard to their general features. One prominent example of tool-transformation is the development of game theory to be used in evolutionary biology.

3. The Migration of Game Theory From Social Sciences to Biology

In this section, I show in what sense the novel use of game theory in evolutionary biology, which is now known as 'evolutionary game theory' ('EGT') can be considered as a tool-transformation. I should mention that my account of the migration of game theory in this paper is not meant to address all the limitations of both game theory and EGT in their respective disciplinary contexts. Instead, the purpose of this account is to show that one *can* detect patterns of migration that have epistemic implications by focusing on the target profile and usage profile of a research tool.

3.1 Game theory in social sciences

Game theory was initially formulated to mathematically model strategic interactions between intelligent, rational agents. In game theory, a game is defined as an interaction between two or

more players in which each player's payoff (e.g., profit) is affected by the decisions made by other players. Typically, such a game assumes both *perfect information* and *common knowledge. Perfect information* assumes that all players know the entire structure of the game (all moves and all payouts) as well as all previous moves made by all players in the game (if it is an iterated or multi-move game). *Common knowledge* is the assumption that all players know that all players have perfect information, and that all players know that all players know that all players have perfect information, and so on. That is, *common* knowledge concerns what players know about what other players know. Moreover, the players also recognize that all players are cognizant that all players are rational, i.e., there is common knowledge of the game and of the *unbounded rationality* of all players. As such, all players will act in the way that takes all other players' potential moves into account in order to maximize their odds of winning. In addition to these assumptions regarding the players of a game, the structure of a game, which refers to the combinations of each move and its payout, is usually summarized in a 'payoff matrix.' Typically, an analysis of a game aims to find out its 'solution,' a unique Nash equilibrium (or sometimes equilibria) of the game.

Game theory has been used in economics, as well as other social sciences, to describe, predict, optimize, or explain a variety of human interactions, such as the economic behaviors of firms, markets, and consumers (e.g., Brandenburger and Nalebuff 1995; Casson 1994) military decisions (Haywood 1954) or international politics (e.g., Snidal 1985).

3.2 Game theory in evolutionary biology

Game theory was later used in evolutionary biology, where a game is understood as phenotypes (or heritable traits) in contest. In 1973, John Maynard Smith and George Price borrowed the formalism of a payoff matrix from game theory to mathematically model the evolution of phenotype frequencies in a population of organisms (see Grüne-Yanoff 2011). Their modeling method assumed that phenotypes are in contest with other phenotypes in a population of organisms. For instance, in a Hawk-Dove game, the contest is embodied by organisms with the phenotype of being aggressive and other organisms that are peaceful. In such a context, the payoff of a move is interpreted as the reproductive success of the phenotype (i.e., the number of copies it will leave to the next generation). Moreover, while the terminology such as 'game,' 'payoffs' and the formalism of a payoff matrix can be seen in the novel use of game theory in biology, the solution to a game in evolutionary biology is decidedly different from the Nash-equilibrium. An evolutionary game theoretic analysis typically looks for an evolutionarily stable strategy (ESS), i.e., a distribution of phenotypes in a population that is stable.

3.3 Epistemic implications of tool transformation

It is clear that the target profile of game theory is no longer the same between its established use

in social sciences and its novel use in biology. First, none of the assumptions of *perfect information*, *common knowledge*, and *unbounded rationality* in what is now known classical game theory (CGT) remain in the novel use of game theory in biology. Second, the moves in EGT are heritable phenotypes exhibited by a group of organisms instead of acts available to players. Third, the payoffs in EGT are the reproductive success of the heritable traits. In this sense, the three assumptions concerning the players were stripped away from the tool - as a result of uprooting game theory from social sciences, and the *heritability* assumption about the moves as well as Darwinian fitness interpretation of the payoff were introduced to the tool - as a result of re-situating it to evolutionary biology.

Note that the change in the target profile forces a limitation to the usage profile of the migrated tool. For instance, nullifying the *unbounded rationality* assumption concerning the players, EGT can no longer be used to optimize a game, i.e., discovering the rationally optimal strategy, which is a common use of game theory in social sciences. For instance, in the prisoner's dilemma, the Nash-equilibrium is for both players to defect. This solution is often interpreted as a prescription for the game; the players are irrational not to defect. However, in a Hawk-Dove game, the ESS obviously has no such normative use. Because the 'moves' of being an aggressive type or a peaceful type are not 'chosen,' the idea of there being normatively better or worse choice of moves is therefore questionable. Moreover, the organisms are not assumed to be rational. Thus, while the players in the prisoner's dilemma could be said to be irrational for choosing to cooperate, this sense of normativity does not carry over to the evolutionary game theoretic analysis of the Hawk-Dove game. One would be mistaken to say that it is 'irrational' for the doves to be doves. Thus, the change in the target-profile of game theory, especially the stripping away of the *unbounded rationality* assumption, has resulted in how the migrated tool should or should not be used.[1]

Moreover, applying EGT to study social phenomena (e.g., Axelrod 1984) or cultural evolution (e.g., Skyrms 2010) requires a careful re-defining of the terms (such as fitness) so as to avoid misinterpretation. Using EGT in social sciences, which can be considered as a 'homecoming' of the migrated tool, is not uncommon. However, the notion of payoffs in EGT refers to, roughly, the overall biological reproductive success of a group of organisms that exhibit a phenotype. Obviously in a social context, reproductive success of the members of some group is not, very often, the feature of interest. A careful reinterpretation of payoffs is thus needed in every analysis to prevent misleading conclusions.

[1] Of course, a more interesting prescriptive use of the ESS of a Hawk-Dove game might be, for example, to manage ecosystems for optimal predator-prey balance. Nevertheless, it should be noted that a justification for this type of prescriptive use of EGT would require further analysis because it is apparently not be derived from CGT.

To generalize, this example suggests that at least in some cases, a change in the target profile requires a corresponding change in the usage profile, or failure of producing genuine knowledge may follow. So far, I have shown that a solution of an ESS analysis may not be interpreted as an optimization to a Hawk-Dove game. Applying EGT to study social phenomena also requires careful treatment to the notion of payoff. Now if, hypothetically, some researcher were to make either of these two mistakes, his or her novel use of the tool would have been classified as tool-transfer - the novel use changes only the target profile without also changing the usage profile. It suggests that in some cases, tool-transformation may not be as risky as tool-transfer. I will come back to the issue of tool-transfer after some remarks related to the migration of game theory.

4. Contributions of the Tool Migration Analysis

The tool migration typology and its focus on tracking both similarities and differences meets the needs to sharpen discussions concerning inter- or cross-disciplinary use of research tools. Current literature seems to lack a framework to capture important, relational characteristics of the research tools that appear in multiple disciplinary contexts. For instance, 'tool-transformation' captures significant differences in details between CGT and EGT without losing sight of the contextual relationship between the two. In contrast, other terms in the literature, such as 'imports' or 'transfers,' fall short of doing so.

'Imports' signals the importation of research tools from a foreign discipline. In contrast, 'transfers' refers to the use of a scientific model, which was established to study phenomena of one domain, to study phenomena of a different domain. Neither term captures the migration of game theory to biology. As Grüne-Yanoff argues,

> [B]iologists constructed the more sophisticated formal [evolutionary game theoretic] concepts themselves. One could speak of the import of formal concepts only with respect to very basic notions such as strategies or pay-off matrices, and it may be more appropriate to refer to formal inspirations rather than imports or transfers in these contexts. (2011, 392)

Moreover, I have suggested that a change in a tool's target profile without a corresponding change in the tool's usage profile *may* lead to misinterpretation and hence misuse of the tool. If this observation is generalizable, which is debatable, then it follows that cases of tool-transfer are epistemically riskier than cases of tool-transformation. On the other hand, if this observation applies only to some cases, it nevertheless reveals at least two epistemic implications concerning tool migration: 1) when the target profile changes, one must be careful not to draw conclusions that might be natural in the old context but may not make sense within the new context, given the new target, and 2) sometimes a change in target profile can, force a change in usage profile. Potentially failing to recognize when these changes occurred in a migration leads

to risky uses of the migrated tool.

Morgan (2011) has argued that while not all scientific knowledge travels far, those that travel with integrity (i.e., maintaining their content more or less intact during its travels) and travel fruitfully (i.e., finding new users or new functions) are considered to be traveling well. It is relatively easy to quantify the latter feature – one needs to look at just the number of a tool's novel applications. However, determining whether a tool has traveled with integrity is not straightforward. As a starting point, this proposed tool migration framework—especially its distinction between the target profile and the usage profile of a tool—provides a starting point that is crucial for assessing the integrity of a migrated research tool. With this framework, one may discover more patterns of tool migration that impact the epistemic integrity and, consequently, effectiveness of a migrated research tool in a foreign discipline.

5. Conclusion

I have argued that mathematical constructs used in science can be viewed as research tools and their cross-disciplinary novel use as tool migration. I have also argued that making novel use of established tools has its risks, but such an implication is not meant to deter cross-disciplinary sharing of tools. Indeed, certain important breakthroughs in the history of science are due to creative, unconventional, uses of research tools (e.g., the use of Fourier's mathematical treatment of heat to study electrostatics [Thomson 1842] or the use of Faraday's mechanical model of fluid motion to model the electromagnetic field [Maxwell 1861]). Versatile research tools are not rare in science. A framework of tool migration aims to offer not only a useful terminology to characterize the diverse landscape of their versatility but also a groundwork to investigate risky patterns of making novel use of established research tools. Finally, this tool migration approach shows that viewing these constructs as tools whilst acknowledging their representational features (i.e., as captured in their target profile) allows for a balanced understanding of knowledge production - especially those productions that are aided by research tools that have migrated across disciplinary boundaries.

References

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Bailer-Jones, Daniela M. 2009. *Scientific Models in Philosophy of Science*. University of Pittsburgh Press.

Brandenburger, Adam M., and Barry J. Nalebuff. 1995. *The Right Game: Use Game Theory to Shape Strategy*. Harvard Business Review.

Boon, Mieke, and Tarja Knuuttila. 2009. "Models as Epistemic Tools In Engineering Sciences: A Pragmatic Approach." In *Handbook of the Philosophy of Science*, edited by Anthonie Meijers, 687–720. Elsevier B.V.

Casson, Mark. 1994. *The Economics of Business Culture: Game Theory, Transaction Costs, and Economic Performance*. Oxford University Press.

Grüne-Yanoff, Till. 2011. "Models as Products of Interdisciplinary Exchange: Evidence from Evolutionary Game Theory." *Studies in History and Philosophy of Science Part A* 42 (2): 386–97.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Haywood Jr, O. G. 1954. "Military Decision and Game Theory." *Journal of the Operations Research Society of America* 2 (4), 365–85.

Humphreys, Paul. 2002. "Computational Models." *Philosophy of Science* 69 (September): 1–27.

___ 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Knuuttila, Tarja, and Andrea Loettgers. 2014. "Magnets, Spins, and Neurons: The Dissemination of Model Templates across Disciplines." *The Monist* 97 (3). The Oxford University Press: 280–300.

Knuuttila, Tarja, and Andrea Loettgers. 2016. "Model Templates within and between Disciplines: From Magnets to Gases–and Socio-Economic Systems." *European Journal for Philosophy of Science* 6 (3). Springer: 377–400.

Maynard Smith, John, and George Price. 1973. "The Logic of Animal Conflict." *Nature* 246: 15–18.

Maxwell, James Clerk. 1861. "Xxv. on Physical Lines of Force: Part I.–the Theory of Molecular Vortices Applied to Magnetic Phenomena." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21 (139): 161–75.

Morgan, Mary, and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Vol. 52. Cambridge University Press.

Morgan, Mary. 2010. "Travelling Facts." In *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, edited by Peter Howlett and Mary Morgan, 3–39. Cambridge University Press.

___ 2014. "Resituating Knowledge: Generic Strategies and Case Studies." *Philosophy of Science* 81 (5). University of Chicago Press: 1012–24.

Morrison, Margaret. 1999. "Models as Autonomous Agents." In *Models as Mediators: Perspectives on Natural and Social Science*, edited by Mary Morgan and Margaret Morrison, 38–65. Cambridge University Press.

Skyrms, Brian. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press.

Snidal, Duncan. 1985. "The Game Theory of International Politics." *World Politics* 38 (1). Cambridge University Press: 25–57.

Thomson, William. 1842. "On the Uniform Motion of Heat in Homogeneous Solid Bodies and Its Connection with the Mathematical Theory of Electricity." *Cambridge Mathematical Journal* 3 (1842): 71–84.

Van Der Steen, Wim J. 1993. "Towards Disciplinary Disintegration in Biology." *Biology and Philosophy* 8 (3): 259–75.

Seattle, WA; 1-4 November 2018 -447-

Representations are Rate-Distortion Sweet Spots

Manolo Martínez (mail@manolomartinez.net)

Abstract

Information is widely perceived as essential to the study of communication and representation; still, theorists working on these topics often take themselves not to be centrally concerned with "Shannon information", as it is often put, but with some other, sometimes called "semantic" or "nonnatural", kind of information. This perception is wrong. Shannon's theory of information is the only one we need.

I intend to make good on this last assertion by canvassing a fully (Shannon) informational answer to the metasemantic question of what makes something a representation, for a certain important family of cases. This answer and the accompanying theory, which represents a significant departure from the broadly Dretskean philosophical mainstream, will show how a number of threads in the literature on naturalistic metasemantics, aimed at describing the purportedly non-informational ingredients in representation, actually belong in the same coherent, purely information-theoretic picture.

1 Information, Shannonian and Dretskean

In what follows I will use a random variable, S, to encode the state the world is in, and another random variable, M, for signals. How should we characterize the information that values of M (i.e., individual signals) carry about values of S (i.e., individual world states)? The most basic quantity with which information theory records dependence among two random variables is the *mutual information* between them. This quantity being an expected value, Dretske (1981, p. 52f) claims, renders it unsuitable for an analysis of representational status, and it should be substituted by notions that record relations between individual states, S_i, and individual signals, M_j. The basic relation which substitutes mutual information in contemporary Dretskean accounts is that of *making a probabilistic difference* (Scarantino 2015): a signal M_j makes a probabilistic difference to the instantiation of a state S_i iff the following *basic inequality* holds:

$$P(S_i|M_j) \neq P(S_i)$$

Nearly all the accounts of information developed in the recent, and not so recent, philosophical literature on this topic are variations on, and attempts to quantify, this inequality. For illustration, in Skyrms (2010, p. 36) the "information in $[M_j]$ in favor of $[S_i]$" is defined as the *pointwise mutual information* (Also *pmi* henceforth) between

state and signal. There is a direct relation between pmis and the basic inequality: the former are nonzero iff the latter is true.

The running thread connecting most prominent contemporary accounts of information is that all there is to Shannon's information theory, at least for the purposes of investigating the nature of representation, is two quantities: the unconditional probability of states and the probability of states conditional on signals, perhaps rearranged as the logarithm of their ratio, or in some other way. Unsurprisingly, from this it is routinely concluded that there is much more to representation than information. This conclusion is premature: informational content in the Dretskean tradition is not by a long shot all there is to information theory. This should not be taken to imply that information is all there is to representation—for one thing, I believe with teleosemanticists (Millikan 1984; Papineau 1987) that teleofunctions have a role to play in a complete theory of representation—but it does mean that no Dretske-style "semanticized information" needs to be recognized, over and above the quantities studied in information theory proper. I will argue that it also means that some prominent proposals as to ways to bridge the information-representation gap are, in fact, unwittingly appealing to informational structure.

In the following section I review two such proposals. My aim is not to argue against them—they are built upon largely correct insights. I will instead aim at showing that a better informed understanding of information provides a way to incorporating these insights in a unified, purely information-theoretic picture.

2 Bridging Information and Representation

2.1 Many-to-One-to-Many Architectures

The first proposal is that it is not enough that representations carry information; on top of that, they must sit in the right place in a certain cognitive architecture. Sterelny (2003), for example, has argued that the emergence of representations is enabled by two prior evolutionary transitions: from "detection" to "robust tracking", on the one hand; from "narrow-banded" to "broad-banded" behavioral responses, on the other. Robust tracking is in essence a *many-to-one* relation between world state and signal: many sensory inputs give rise to one and the same representation. Other theorists have advocated similar architectural constraints on representational vehicles. Famously, Burge (2010) places a great deal of weight on *perceptual constancies* in his characterization of perceptual representation (Burge 2010, p. 413.) This is a variation on Sterelny's idea and, as such, a many-to-one architectural constraint on representational status.

As for broad-banded responses, in these systems a single representation will be flexibly dealt with, resulting in different courses of action, depending on the context where the representation is tokened. Response breadth is in essence a *one-to-many* relation between representational vehicle and output: one representation, many agential outputs.

2.2 Reference Magnetism

A second proposal has been to focus on the entities that should figure in the content of simple representations. The suggestion, typically, is that represented entities should be appropriately *natural*, or *real*. For example, Dan Ryder (2004, 2006) has argued that neurons become attuned to *sources of correlation*. These entities are closely related to Richard Boyd's *homeostatic property clusters* (also HPC henceforth, Boyd 1989): HPC theory identifies natural kinds with clusters of properties which tend to be instantiated together, and such that this frequent co-instantiation is not just a statistical fluke. What Ryder calls sources of correlation are the grounds for these HPC-related frequent co-instantiations—whatever it is that makes them *not* statistical flukes. Ryder claims that many of the representations the brain trades in target sources of correlation. Martínez (2013) and Artiga (forthcoming) have made more general cases that simple representations preferably target HPCs (Martínez), or properties that best explain the co-occurrence of other properties (Artiga).

A similar idea has been explored in an entirely independent line of enquiry starting with Lewis (1983): "among the countless things and classes there are ... [o]nly an elite minority are carved at the joints, so that their boundaries are established by objetive sameness and difference in nature. Only these elite things and classes are eligible to serve as referents" (Lewis 1984, p. 227). This is what Sider (2014, p. 33) calls *reference magnetism*.

As I show in section 4, these two ideas, although apparently disparate, are in fact closely related, and the explanatory payback they bring to representation-involving talk depends on their informational underpinnings.

3 Information Theory is a Source-Channel Theory

Philosophy has understood information theory as a mostly *definitional* effort: for all philosophers have typically cared, the theory begins and ends with a presentation of what it takes for one random variable (or the worldly feature it models) to carry information about another. But information theory goes well beyond that. It is, well, a *theory*, and as such it is chiefly composed of claims that are advanced in the hope that they be true about the world.

In a nutshell, the most celebrated results in information theory have to do with specifying how faithful the transmission of information from a source can be, when it happens over a (typically noisy, typically narrow) channel. These results have played absolutely no role in informational accounts of representation.[1] Take, for starters, the idealized depiction of an information-processing pipeline in fig. 1 (*cf.* Cover & Thomas 2006, fig. 7.1)

[1] Two recent philosophical treatments of information that try to redress this neglect are Mann (2018) and Rathkopf (2017).

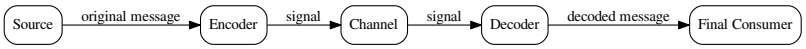


Figure 1: An information-processing pipeline

Here an *encoder* produces a signal as a response to information incoming from a source. This signal goes through a channel and is subsequently decoded, producing a message that is then utilized for whatever purposes downstream. The first thing to note is that the broadly Dretskean ideas about the content of a signal introduced in section 1 only have use for the first two links in this information-processing chain: how signals carry information about a certain original message produced by a source, as depicted in fig. 2. In fact, in information theory the main action happens immediately after that: a source is producing stuff, and we want that stuff to *go through a channel*. Information theory is mainly about providing theoretical guarantees of faithfulness in transmission, given the rate of the channel. We can think of this rate as the number of bits it provides for the encoder to use in the signal. If, say, the rate is 2 bits per use of the channel, this means the encoder can use up to 2 bits to construct the signal and be sure that it can pass unscathed through the channel and on to the decoder.

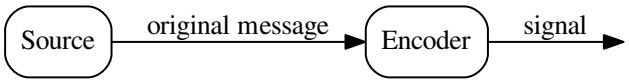


Figure 2: The information-processing pipeline in the Dretskean tradition

In typical cases of representation, channel rate is consistently smaller than ideal. Consider animal alarm calls. Vervet monkeys, for example, are typically described as being able to produce three different, discrete kinds of calls (Seyfarth, Cheney & Marler 1980a, 1980b) that are usually taken to be associated with the presence of leopards, eagles and snakes respectively. Obviously, the entropy of the relevant aspects of the environment that prompt the production of a call (think of all the possible patterns of approach of these predators, for example) vastly outstrip the rate of a channel, which consists in the production of just one out of three possible signals. This means that loss in communication is inevitable. Alarm calls, and for analogous reasons representations in general, are all about *lossy transmission*.

The way in which information theory deals with lossy transmission is by defining a *distortion measure* (Cover & Thomas 2006, p. 304) that gives a score to a pair composed of a certain original message M, and the decoded version thereof, $\hat{M}$. In what follows I

will be using the *Hamming distortion* which simply adds 1 to the distortion when the bits in the original and decoded signals (which we can assume to be binary strings) do not coincide, and 0 otherwise, then normalizes. So, for example, the Hamming distortion between an original signal $M = 010011$ and a decoded signal $\hat{M} = 100010$ is ⅓, because the first, second, and last (a total of 3) bits have been decoded incorrectly, and there are 6 bits in total.

The central result in this so-called *rate-distortion theory* approach to lossy transmission is that there is a *rate-distortion function*, $R(D)$, which gives the minimum rate at which any given distortion is achievable. The actual mathematical expression of the rate-distortion function need not detain us here (see Cover & Thomas 2006, p. 307, theorem 10.2.1), but it is such that the *Blahut-Arimoto* algorithm (Blahut 1972; Arimoto 1972) allows us to calculate it easily.

The main thesis of this paper is that representations belong in information-processing pipelines whose rate-distortion function has *sweet spots*: by this I mean points in the rate-distortion curve such that the usefulness of increasing the rate of the channel past those points is much smaller than before reaching them. Moreover, the encoding-decoding strategies that make use of these representations tend to live in the vicinity of those sweet spots. I submit that it is these information-theoretic properties that the conditions on representation discussed in section 2 try to get at.

To see how rate-distortion analyses work let's start by looking into a source that models a series of fair-coin tosses: this random variable would have two values, *heads* and *tails*, with associated probabilities $P_{heads} = P_{tails} = .5$). Using the Hamming distortion as our target distortion measure, if the coin lands heads (tails) and the decoded message is tails (heads) the distortion is 1, otherwise 0. The Blahut-Arimoto algorithm allows us to draw the rate-disortion curve, in fig. 3. Here the blue line is the rate-distortion curve. It intersects the x-axis at 1.0 bits (the entropy of the source) and it intersects the y-axis at 0.5 (the lowest average distortion one can achieve when the channel is closed.) The red line gives a measure of how steep the blue line is at any given point—in particular, the absolute value of the slope of the blue line. The higher the red line, the steeper the blue line.

The situation this setup is modeling is one in which a single cue is present or absent, and a signal tries to keep track of whether it does. This is precisely the kind of situation where many theorists (certainly Sterelny and Burge, for the reasons reviewed in 2.1) would see the postulation of representations as entirely idle—see, e.g., Schulte's vasopressin example in his Schulte (2015). In agreement with the idea that postulating representations here is idle, there is not much structure to the rate-distortion curve corresponding to this setup: reading the chart from right to left, increasing the rate makes the achievable expected distortion go smoothly down, until the rate hits the entropy of the source, at which point the achievable distortion is zero.That's about it.

Let's now model one kind of situation in which there is a reasonably wide consensus that representations make an explanatory contribution: vervet-monkey alarm calls, as reviewed above. In the model, the source—the situation the information-processing pipeline is dealing with—randomly makes members of two natural kinds (we can think

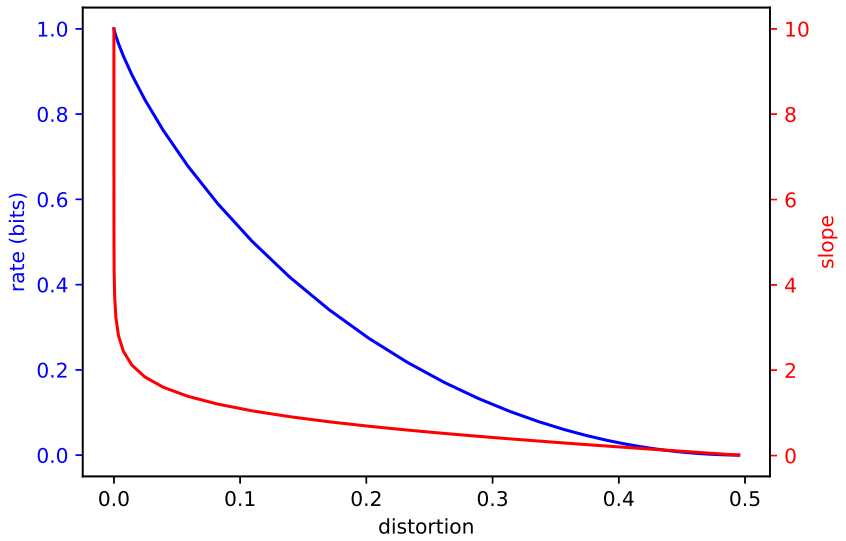


Figure 3: The rate-distortion function for a coin toss

of them as two different predators) be or not present at any given time, independently from one another. This intends to mimic the situation vervet monkeys face, where snakes, leopards and eagles show up or not, more or less at random.

These natural kinds are modeled as homeostatic property clusters (see section 2.2 above). In order to derive a explicit probability distribution for the source out of this qualitative description, the two HPCs are in their turn represented by two Bayesian networks, each with a parent node and four children (see fig. 4.) Each of the nodes stands for a property; if the node is *on* it means the corresponding property is instantiated; if it is *off* it means it is not. In the model, children nodes replicate noisily the state of their parent. Thus, e.g., if the parent is *on* (if the corresponding property is instantiated) each child property will have a .95 chance of being instantiated too; if the parent is *off* the probability for each children of being instantiated is .05. The unconditional probability of instantiation for the two parent nodes is .5.

In the model, the source produces a binary string, with each member of the string being 1 if the corresponding node is on, and 0 if it's off. This signal is encoded, goes through a channel, and is then decoded at the other side. The target distortion measure is the Hamming distortion. Fig. 5 plots the rate-distortion curve for this model.

This curve is very different from the one in fig. 3: there is a clear "sweet spot"—a sudden drop in the usefulness of extra rate, see the red curve—when the system hits a rate of 2 bit/use. I.e, there is, in a certain principled sense, an optimal level of lossy compression; a way to set up an encoding-decoding strategy that recover most of what's going on in

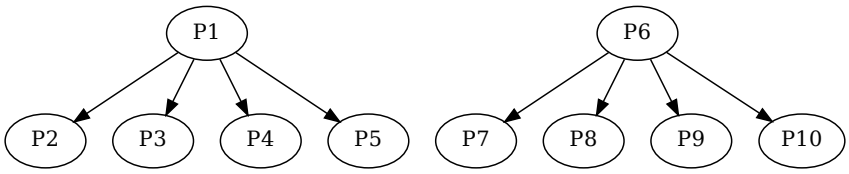


Figure 4: Two natural kinds

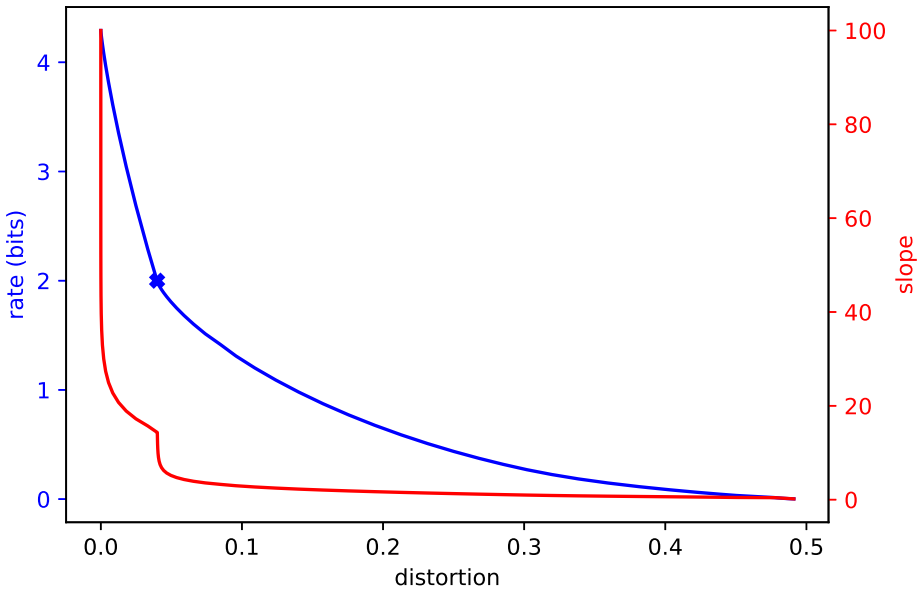


Figure 5: A sweet spot in the rate-distortion function

the world of relevance to the information-processing system, even through a very severe, 2 bit bottleneck. I claim that this is no coincidence. Our representation-attributing practices gravitate towards this kind of situations.

To see how sweet spots in rate-distortion curves and representations are related, consider now what an optimal encoding-decoding strategy would look like. That is, how should the encoder encode the information coming from the source, and how should the decoder decode the signal coming from the encoder, so that the resulting expected distortion between original and decoded signal is the minimum achievable, at the sweet spot?

Optimal Encoding Strategy: First divide the incoming signal in two halves, one corresponding to properties P_1 through P_5; the other corresponding to properties P_6 through P_{10}.

If there is a majority of 1s in the first half of the original signal set the first bit of the signal to 1. Otherwise set it to 0. Ditto for the second half of the original signal and the second bit of the signal.

Optimal Decoding Strategy: If the first bit in the incoming signal is 1, set the first half of the decoded signal to 11111. Otherwise, set it to 00000. Ditto for the second bit and the second half of the decoded signal.

How should we interpret what encoder and decoder are doing here? A natural way is this: they are using the presence or absence of properties in an HPC cluster as diagnostic of the presence or absence of the underlying natural kind—this would be the encoding part—and then taking the resulting signals as representing the presence of a paradigmatic instance of the kind, one that has all the properties in the cluster—this would be the decoding part. HPC kinds being what they are, frequently the first half of the incoming signal will resemble the paradigmatic presence of the first kind (11111) or its paradigmatic absence (00000), and the same will happen with the second half and the second kind. That is why this encoding-decoding strategy works so well.

In describing this optimal strategy I have helped myself to representational vocabulary; it has been useful in order to explain how the strategy works, and how come that behaving in this particular way achieves low distortion at low rates: it is because each of the two bits in the signal is caused by, and causes, behavior that is optimally attuned to the probabilistic structure of each of the two natural kinds in the model world, respectively. Nothing going on in this system falls outside the purview of Shannonian information theory—of information theory *tout court*, so at least in this kind of cases representational talk depends on no non-informational fact.

We can now understand better what's lacking in the philosopher of mind's information-theoretic tookit: it is entirely possible, and computationally trivial, to calculate, e.g., Skyrms's pmi between each of the possible signals (00, 01, 10 and 11) and each of the possible world states (all 1024 of them, from 0000000000 to 1111111111). Doing so would leave us with 4 vectors (one for each signal) with 1024 entries each (one for each world state.) First, this is an unwieldy collection of numbers, which doesn't bring out the relevant structure. For example, if the probability of children nodes being *on* conditional on their parent being *on* was .96 instead of .95 the rate-distortion curve

would be qualitatively identical, with a sweet spot in exactly the same place, yet most numbers in the Skyrmsian informational content vectors would change. Second, and most important, nothing in those 4096 numbers allows us to infer the presence of a sweet spot. The relevant information is simply not there, depending as it does on a distortion measure which is not used in computing Skyrmsian informational contents.

If this is approximately right, the question about what makes representational talk explanatory is readily answered: saying that a certain vehicle is a representation conveys something quite specific about its informational context. It says that the vehicle is part of an encoding-decoding strategy that exploits a sweet spot in a rate-distortion curve—where the curve is in turn fixed by the probabilistic structure of the world, and the target distortion measure. This, in less technical terms, translates to saying that the vehicle is summarizing *relevant* (this is where the distortion measure comes in) aspects of the current situation in an optimal, if lossy, manner, made possible by *how the world* is (this is where the probabilistic structure of the world comes in.) This explication of the explanatory contribution of representations can be turned into an explicit answer to what makes something a representation—an answer, that is, to what Artiga (2016) calls the metasemantic question.

The Rate-Distortion Approach: A signal, S, in a certain information-processing pipeline, P, is a representation if the following two conditions are met:

Existence: There are sweet spots in the rate-distortion curve associated with P.

Optimality: S is produced as part of an encoder-decoder strategy that occupies the vicinity of one of these sweet spots.

So, *pace* Dretske, the core information-theoretic notions of entropy, rate, distortion, etc. *can* provide invaluable insight into the representational status of individual signals. If the rate-distortion approach is on the right track, those information-theoretic notions, through the existence condition, specify the kind of setup where representations live, which then the optimality condition can use to provide a criterion for the representational status of individual signals.

I offer the foregoing discussion as a preliminary case for the rate-distortion approach to representation: it shows how postulating representations is explanatory, even if these representations depend just on (Shannon) information. It illuminates the difference in representational status between cue-driven examples, such as Schulte's vasopressin; and vervet alarm calls, and other similar examples. To complete my case I now show how the ways to bridge the gap between natural and nonnatural information discussed in section 2 can be seen as unwitting attempts to get at rate-distortion sweet spots.

4 There is no Gap to Bridge

What does it take for the existence condition to be met? That is to say, what circumstances result in sudden drops in the slope of the rate-distortion curve? We have seen one such family of circumstances: if the pattern in which properties are instantiated

in the source is noisily replicated in a cluster then sudden drops are to be expected: distortion will decrease with rate up to the point where all the main sources of variation in property instantiations are accounted for, and all that remains is the residual noise in instantiations within each cluster. Take a look again at figs. 4 and 5: to describe this source we basically need enough rate to account for the two main sources of variation: P_1 and P_6. This is not all there is to the world, because it's possible for the other properties to (fail to) token independently of their parent, but the unlikeliness of these departures makes the extra rate comparatively less useful.

Noisy replication of property instantiations is at the core of the HPC theory of natural kinds, as we saw above. This means that, in general, the presence of HPC natural kinds in a source will create sweet spots. This opens a line of argument in favor of reference magnetism from information-theoretic premises: reference magneticsm should be seen as making a point about the kind of probabilistic structure that an information-processing pipeline must be attuned to, if signals are to effect the kind of optimal lossy compression that underlies our representation-attributing practices. Reference magnetism is just a way of meeting part of the existence condition.

Regarding the suggestion, by Sterelny, Burge and others, that representations inhere preferably on signals sitting in a one-to-many-to-one pipeline, I submit that the many-to-one aspect of this suggestion aims at meeting the optimality condition; the one-to-many aspect, together with reference magnetism, aims at meeting the existence condition.

The first thing to note here is that the *Optimal Encoding Strategy* presented above enforces what Sterelny calls robust tracking and Burge calls constancy: the strategy consists in considering all properties coming from each of the two clusters and setting the relevant bit to 1 only if a majority of those properties are instantiated. That is, the encoder is taking a multiplicity of configurations (e.g., the first half of the incoming signal being 00111, 01011, 10111, etc.) to a single output: the first bit of the signal being 1. Furthermore, that part of the signal will be decoded as 11111: from there on, the system downstream will treat whatever is out there in the world as a paradigmatic member of the first kind. The system is recovering the presence of a natural kind out of many different, noisy instantiation patterns. This is a clear instance of constancy. Suppose that the encoder, insted of being many-to-one, depended on a single cue; say, suppose it set the first bit to 1 if one of the children properties (say, P_2) was instantiated, and to 0 otherwise. In such a cue-driven setup, the best encoder-decoder arrangement possible is marked by the blue circle in fig. 6. This has double the distortion than than the optimal encoding (marked by the blue cross) which sits right on top of the optimal rate-distortion curve. This cue-driven system would not meet the optimality condition, which means that a many-to-one architecture is instrumental to meeting it.

Finally, the target distortion measure in the information-processing pipeline can be seen as that which Sterelny's one-to-many condition on representation is actually tracking. Using, for example, the Hamming distance as a distortion measure is tantamount to assuming that all of the properties of the natural kinds are relevant for downstream processing. One natural way in which this may happen is when the agent is to respond flexibly to the presence of the natural kind: in different contexts or states different properties of the kind might be relevant and, for example, the presence of a tree might

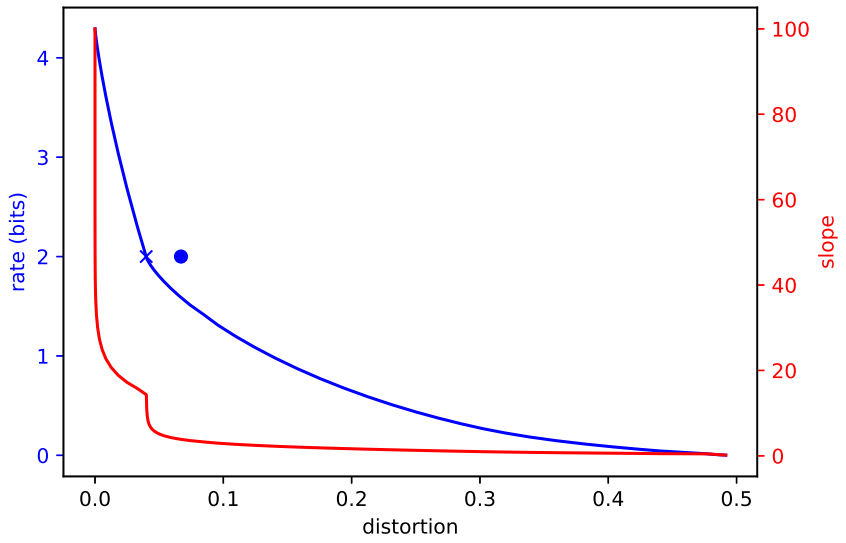


Figure 6: Cue-driven encoding

be sometimes relevant to behavior because it bears fruit (if the agent is hungry) and some other times because it has a dense cover (if the agent is looking for shelter.)

Caring about all (or many) properties of the kind is what makes the rate-distortion curve display a sweet spot. If, instead, the agent has a rigid, stereotyped response to the presence of members of the kinds—that is, if it only cares about the presence of one property, which is the property that makes that rigid behavioral response fitness-conducive, then the curve is as presented in fig. 7. Rigid behavioral responses make the probabilistic structure of the kinds largely irrelevant. As a result, the system behaves as if a coin were tossed, where heads would mean that the target property is tokened, and tails that it is not. This arrangement does not meet the existence condition. Sterelny's broad-banded responses are, again, a way of getting at rate-distortion sweet spots.

References

Arimoto, S 1972, 'An algorithm for computing the capacity of arbitrary discrete memoryless channels', *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20.

Artiga, M forthcoming, 'Beyond Black Spots and Nutritious Things: A Solution to the Indeterminacy Problem', *Dialectica*.

Artiga, M 2016, 'Liberal Representationalism: A Deflationist Defense', *dialectica*, vol.

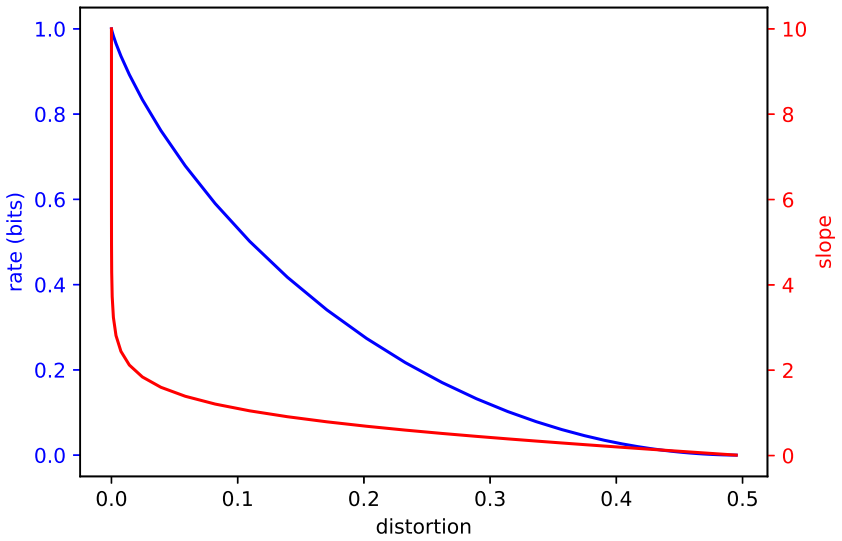


Figure 7: Rigid behavioral response

70, no. 3, pp. 407–430.

Blahut, R 1972, 'Computation of channel capacity and rate-distortion functions', *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473.

Boyd, R 1989, 'What Realism Implies and What It Does Not', *Dialectica*, vol. 43, no. 1-2, pp. 5–29.

Burge, T 2010, *Origins of objectivity*, Oxford University Press.

Cover, TM & Thomas, JA 2006, *Elements of Information Theory*, New York: Wiley.

Dretske, F 1981, *Knowledge and the Flow of Information*, The MIT Press.

Lewis, D 1983, 'New work for a theory of universals', *Australasian journal of Philosophy*, vol. 61, no. 4, pp. 343–377.

Lewis, D 1984, 'Putnam's paradox', *Australasian Journal of Philosophy*, vol. 62, no. 3, pp. 221–236.

Mann, SF 2018, 'Consequences of a Functional Account of Information', *Review of Philosophy and Psychology*, pp. 1–19.

Martínez, M 2013, 'Teleosemantics and Indeterminacy', *Dialectica*, vol. 67, no. 4, pp.

427–453.

Millikan, R 1984, *Language, Thought and Other Biological Categories*, The MIT Press.

Papineau, D 1987, *Reality and Representation*, Basil Blackwell.

Rathkopf, C 2017, 'Neural information and the problem of objectivity', *Biology & Philosophy*, vol. 32, no. 3, pp. 321–336.

Ryder, D 2006, 'On Thinking of Kinds', in G Macdonald & D Papineau (eds), *Teleosemantics*, Oxford University Press, pp. 1–22.

Ryder, D 2004, 'SINBAD Neurosemantics: A Theory of Mental Representation', *Mind & Language*, vol. 19, no. 2, pp. 211–240.

Scarantino, A 2015, 'Information as a probabilistic difference maker', *Australasian Journal of Philosophy*, vol. 93, no. 3, pp. 419–443.

Schulte, P 2015, 'Perceptual representations: A teleosemantic answer to the breadth-of-application problem', *Biology & Philosophy*, vol. 30, no. 1, pp. 119–136.

Seyfarth, RM, Cheney, DL & Marler, P 1980a, 'Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication', *Science*, vol. 210, no. 4471, pp. 801–803.

Seyfarth, RM, Cheney, DL & Marler, P 1980b, 'Vervet monkey alarm calls: Semantic communication in a free-ranging primate', *Animal Behaviour*, vol. 28, no. 4, pp. 1070–1094.

Sider, T 2014, *Writing the Book of the World*, Reprint edition., Oxford University Press, Oxford.

Skyrms, B 2010, *Signals: Evolution, Learning & Information*, New York: Oxford University Press.

Sterelny, K 2003, *Thought In A Hostile World: The Evolution of Human Cognition*, John Wiley & Sons, Malden, MA.

The Proportionality of Common Sense Causal Claims

Jennifer McDonald

This paper defends strong proportionality against what I take to be its principal objection – that proportionality fails to preserve common sense causal intuitions – by articulating independently plausible constraints on representing causal situations. I first assume the interventionist formulation of proportionality, following Woodward.[1] This views proportionality as a relational constraint on variable selection in causal modeling that requires that changes in the cause variable line up with those in the effect variable. I then argue that the principal objection derives from a failure to recognize two constraints on variable selection presupposed by interventionism: *exhaustivity* and *exclusivity*.

[1] Woodward 2003

1. Introduction

Yablo's principle of proportionality holds, roughly, that something counts as a cause of some effect just in case it includes the appropriate degree of causal information.[2] Proportionality has been put to various philosophical uses, such as a proposed solution for the causal exclusion argument, and as a justification and explanation of the dependence on high-level causal explanations in the special sciences. However, the precise formulation of such a principle has proven to be controversial.

I take the most promising formulation to be an interventionist one, following Woodward.[3] Such a formulation defines proportionality as a relational constraint on variable selection in causal modeling. In this paper, I argue that this formulation works well as it is – contra Franklin-Hall (see 2016) – so long as we recognize two independently plausible background requirements on variable selection. I call these *exhaustivity* and *exclusivity*. Exhaustivity holds that a variable must take at least one of its values. Exclusivity holds that a variable can take at most one of its values. Both constraints are relative to, and thereby help to make explicit, the modal assumptions implicit in causal inquiry.

Finally, with these requirements in place, I defend proportionality against its principal objection: that it fails to preserve fundamental causal intuitions. I demonstrate how this concern derives from a failure to recognize and integrate the modal assumptions implicit in causal inquiry, in tandem with an inappropriate use of variables to represent causal situations.

2. Interventionism

The formulation of proportionality that I endorse comes directly from Woodward, and is defined in terms of his interventionist account of causation. Interventionism expands on the intuition that causal claims provide

[2] Yablo 1992
[3] Woodward 2003, 2008a, 2008b, 2010, 2016

manipulability information. If *X* causes *Y*, then manipulating or changing *X* is a way of manipulating *Y*. It then exploits the language of causal models to identify and articulate different causal relations of interest. A causal model can take a variety of forms, such as graphical, potential-outcome, and structural-equations models.[4] However, I'll restrict discussion of causal models in this paper to graphical models. A graphical model is, essentially, a set of variables – representing the causal relata – and a directed binary relation between them – representing causal influence.

Interventionism then defines the notion of an *intervention* on a system. An intervention, *I*, first must directly change the value of some variable, *X*, in such a way that it breaks the dependence that *X* may have had on other variables in the system. Second, *I* must be designed in such a way that any change in the effect variable, *Y*, will be the direct result of *X* and not of *I* itself. Finally, *I* must be wholly independent of other possible causes of *Y*, whether such causes are represented by the given model or not. A more precise formulation than this won't matter for the purposes of this paper.[5]

With this in place, the interventionist then defines a basic notion of cause, which corresponds most closely with the intuitive notion of *causal relevance*:

> (Principle M) *X* causes *Y* iff there are background circumstances B such that if some (single) intervention that changes the value of *X* (and no other variable) were to occur in B, then Y would change. (Woodward 2003, 222)

That is, in order for *X* to be a cause of *Y*, the change in *X* from one value to another as the result of an intervention corresponds to the change in *Y* from one value to another, given some fixed set of background parameters. Various kinds of causal relations are then captured by refinements on this basic notion. Due to

[4] See Greenland and Brumback 2002 and Hitchcock 2009 for overviews of causal models.
[5] See Woodward 2003, chapter 3, especially 98

the irrelevance of these and further details to my argument, I'll leave my overview of interventionism here.[6]

3. Proportionality as Relational Constraint on Variable Selection

Interventionism places variables front and center in how we represent and inquire into causation. Thus, more needs to be said about the criteria for variable selection. Although the variables can be taken to represent different things, I will assume throughout that the set of values of a particular variable represents a set of properties – constrained by a given property type – that are possibly instantiated by some particular thing. The assumed causal relata of this paper will therefore be property instantiations.

This paper addresses two questions relevant to variable selection: (i) What determines the range of values that a variable can take? (ii) At what level of description should the values of the variables be? Proportionality has been proposed as an answer to (ii). However, after laying out the proposal, I'll go on to argue that while (ii) can be answered by the principle of proportionality, it can only do so alongside an appropriate answer to (i). One aspect of such an answer is that the background modal context determines the range of values that a variable takes.

Constraints on variable selection can be divided into two kinds: relational constraints and non-relational constraints. *Relational constraints* pertain to the extrinsic nature of the variables in a causal model, to how "variables relate to one another." (Woodward 2016, 1056) One example of such a constraint is stability.[7] *Stability* is the persistence of the causal relation between a cause variable and an effect variable, despite changes in the background conditions. The more changes such a relation can survive, the more stable it is.

[6] See Woodward 2003, chapter 2, especially section 3
[7] See Woodward 2010, 2016

Proportionality is just such a relational constraint. It holds that changes in a cause variable should line up with changes in an effect variable. Intuitively,

> Proportionality has to do with whether changes in the state of the cause 'line up' in the right way with changes in the state of the effect and with whether the cause and effect are characterized in a way that contains irrelevant detail. (Woodward 2010, 287)

Take Yablo's pigeon example.[8] Sophie the pigeon is trained to peck at red things and only at red things. She then pecks at a paint chip, which is a particular shade of red – scarlet. Which of the following is causally relevant to Sophie's pecking: the chip's being red or the chip's being scarlet?

When translated into interventionist terms, this becomes a false dichotomy. Take the variable, *P*, to be a variable representing whether the pigeon pecks or not. It can take the values: {*peck*, *not-peck*}. Now consider two alternative variables for representing the property-instantiations of the paint chip: the variable, *R*, which can take the values {*red*, *not-red*}, and the variable, *T*, which can take the values {*taupe, scarlet, cyan, mauve, crimson,* etc.}, where 'etc.' stands for all other physically possible colors at the same grain as those already made explicit. According to Principle M, the causal model in which *R* stands as causally relevant to *P* is just as accurate as one in which *T* so stands. In the *R* model, *R* is causally relevant to *P* because an intervention on *R* that changes its value from *not-red* to *red* changes *P*'s value from *not-peck* to *peck.* In the *T* model, *T* is causally relevant to *P* because an intervention on *T* that changes its value from *taupe* to *scarlet* changes *P*'s value from *not-peck* to *peck*.

Interventionism therefore doesn't ask the question, which variable stands in a causal relation to *P*? For, the answer is 'both'. *R* and *T* are each causally relevant to *P*. But, this doesn't mean that their respective relationship to *P* is the same. *R* is *proportional* to *P*, while *T* is not. All of the changes in *R* line up with changes in *P* – every intervention on *R* corresponds to a change in *P*. But only some of the

[8] Yablo 1992

changes in *T* line up with those in *P* – only certain interventions on *T* correspond to changes in *P*. The intervention that changes the value of *T* from *taupe* to *cyan*, for example, will not change the value of *P*.

Woodward defines proportionality more explicitly as,

> (P) There is a pattern of systematic counterfactual dependence (with the dependence understood along interventionist lines) between different possible states of the cause and the different possible states of the effect, where this pattern of dependence at least approximates to the following ideal: [it] should be such that (a) it explicitly or implicitly conveys accurate information about the conditions under which alternative states of the effect will be realized and (b) it conveys only such information – that is, the cause is not characterized in such a way that alternative states of it fail to be associated with changes in the effect. (2010, 298)

There are two views on what this difference between variables like *R* and *T* means. The first takes proportional variables such as *R* to represent genuine causes, while non-proportional variables such as *T* represent merely causally relevant factors. Proportionality is thereby considered a necessary constraint on causation. Call this *strong proportionality*.[9] The second view takes proportionality to be a merely pragmatic constraint on causal explanation.[10] Call this *weak proportionality*. Throughout this paper, I assume and defend strong proportionality.

4. Non-Relational Constraints: Exhaustivity and Exclusivity

Non-relational constraints, on the other hand, pertain to the intrinsic nature of the variables in a causal model. These constraints "can be applied to variables, individually, independently of how they relate to other variables." (Woodward

[9] See List and Menzies 2009; Menzies and List 2010; and Papineau 2013
[10] See Woodward 2015; Shapiro and Sober 2012; McDonnell 2017; and Weslake 2013, 2017

2016, 1057) One example is *metaphysical naturalness*, which requires that variables pick out only natural properties, on some understanding of 'natural'.[11]

What I propose to call the exhaustivity and the exclusivity constraint are similarly non-relational constraints. Take exhaustivity first. The *exhaustivity constraint* requires that a variable's values capture the entire range of relevant possibilities for whatever type of thing the variable represents. An exhaustive variable is one that must take one of its values, given whatever background modal constraints are in place.

Since I've restricted this discussion to variables whose values represent the property instantiation of some target object, I can define exhaustivity in more precise terms. *Exhaustivity* is the constraint on a variable in a causal model that holds that its values must jointly represent the range of possibilities of property instantiation by the given object for the given property-type. If the property-type is a color, for example, then the values must somehow exhaust the color spectrum. This can be done quite simply with a binary variable that can take the values: {*some particular color*, *not-(that particular color)*}.

Next, the *exclusivity constraint* holds that the values of a given variable should be such that any one excludes all the others. Woodward references exclusivity when he writes,

> When considering the values of a single variable, we want those values to be logically exclusive, in the sense that variable *X*'s taking value *v* excludes *X*'s also taking value *v'*, where $v \neq v'$. (2016, 1064)

In other words, if two things are not exclusive – if they could occur together – then they should be represented by distinct variables. While exhaustivity holds that a variable should take *at least* one of its values, exclusivity holds that a variable should take *at most* one of its values.

[11] See Lewis 1983; Menzies 1996; Paul 2000; and Franklin-Hall 2016

Importantly, exhaustivity and exclusivity are each relative to a background modal context. In possible worlds terminology, the modal context is the set of possible worlds relevant to the truth of the counterfactual that captures the causal claim. It can be described as a set of worlds, or perhaps more succinctly as a list of background assumptions that define such a set. These assumptions can include any constraint that operates in a law-like fashion.

For example, the causal claim, "The chip's being scarlet caused the pigeon to peck," corresponds to the counterfactual, "Had the chip not been scarlet, the pigeon wouldn't have pecked." The modal context of this claim and corresponding counterfactual is the set of possible worlds that determines whether the counterfactual is true. So, if this claim and counterfactual are meant to represent a *specific* causal situation near a local paint chip factory that specializes in just the colors scarlet and cyan, and no others, then the relevant set of possible worlds will be constrained to those in which the paint chip takes one of the two factory colors – cyan or scarlet. In this context, the variable, *C,* that can take the values {*cyan, scarlet*}, is an exhaustive variable. Further, given this set of worlds, the counterfactual is true.

If instead these are meant to represent any *general* causal situation involving paint chips and a red-pecking pigeon, then the relevant set of possible worlds will be more inclusive, including all worlds in which the paint chip takes any color within the color spectrum. *C* is not exhaustive relative to this more inclusive modal context. But the variable *T*, from before, is. Given this more inclusive set of worlds, the counterfactual is false, since the pigeon will peck in response to shades of red other than scarlet.

A point of note here is that the constraints of exhaustivity and exclusivity are indeed non-relational constraints in the sense previously defined. Although they are relative to the modal context, they are *not* relative to other variables in the model. They are properties of a variable taken independently as a representation of the target scenario.

I hold that causal models successfully represent causal situations in part by requiring exhaustive and exclusive variables. Proportionality, defined in terms of causal models, also requires exhaustive and exclusive variables. A significant upshot of this is that the proportional cause is not only relative to the target effect variable, but also to the background modal context.

5. Interventionist Proportionality Does the Trick

Franklin-Hall contends that Woodward's formulation of proportionality doesn't successfully prioritize intuitively proportional causal relata, such as red in the pigeon example. However, as I'll argue, presupposing my notion of exhaustivity corrects for this objection.

Franklin-Hall argues that proportionality as laid out in section 3 is inadequate for capturing the kind of causal explanation we're looking for. To do so, she calls upon Sophie and her paint chip. She then introduces a comparison between the causal variable, *R*, that can take the values: {*red*, *not-red*}, (as above), and a variable, *C*, that can instead take the values: {*cyan, scarlet*} (as above). *R*, as before, is proportional to, and therefore a genuine cause of, *Y*. But, she argues, *C*, too, is proportional to *Y*, since every possible intervention on *C* changes the value of *Y*. An intervention on *C* that changes its value from *cyan* to *scarlet* changes *Y* from *not-peck* to *peck*, and an intervention that changes *C*'s value from *scarlet* to *cyan* changes *Y*'s value from *peck* to *not-peck*. Thus, the changes in *C* line up with the changes in *Y* just as well as the changes in *R* do. The problem, then, is that proportionality, as formulated, is insufficient to its intended task. It fails to privilege a variable like *R* over one like *C*, and so fails to prioritize a causal model that uses *R* over one that uses *C*.

In response to this problem, a natural move would be to find a way to disqualify variables like *C* from the arena. Intuitively, *C* is not the right kind of variable. But, why not? I propose that our aversion to variables like *C* is due to their failure to exhaustively represent the implicit modal context of the situation. The background possibilities relative to the paint chip include the full color spectrum.

Unless the possible color of the paint chip is restricted in some way – by the local factory, for example – then the target object can fail to take one of *C*'s two values. There are other physically possible colors that the paint chip could have – such as beige or olive green – and *C*'s values fail to represent these possibilities.

Relative to the implicit modal context, then, *C* is not an exhaustive variable. The variable, *R*, on the other hand, is exhaustive, since the object must take one of *R*'s two values. By requiring exhaustive variables, *C* is discounted as a candidate variable *relative to the implicit modal context*, and *R* takes privilege as the proportional cause.

In general, two variables are in proper competition with each other over which is proportional to some effect variable only when they are exhaustive relative to the same modal context. *C* and *R* are not competitors for proportionality relative to *Y*, since only one of them can contain an exhaustive set of active possibilities relative to any given modal context.

6. Preserving Causal Intuitions

The strongest objection to proportionality, as raised by Bontly, Shapiro and Sober, McDonnell, and Weslake, is that it seems to render many common sense causal claims false.[12] Call this the *objection from common sense*. It objects to strong proportionality by attempting to demonstrate that if proportionality is required of something to be a cause, then many things that we would naturally call causes don't actually qualify.

Take as an example the situation where Socrates drinks hemlock and then dies, and the corresponding causal claim, 'Socrates's drinking hemlock caused him to die'. The objection goes that drinking hemlock is not actually proportional to Socrates dying. For example, if Socrates had not drank hemlock, but still consumed it – by eating a dozen leaves, for example – then he still would have

[12] See Bontly 2005; Shapiro and Sober 2012; McDonnell 2017; and Weslake 2013, 2017

died. This seems to show that the changes in the variable that represents Socrates drinking hemlock don't line up with the changes in the variable that represents Socrates dying. The first variable could change values from *Socrates-drinks-hemlock* to *Socrates-eats-hemlock* and the second variable would retain the value *Socrates-dies*. This common sense causal claim is therefore not proportional. The proportional cause should be, instead, *consuming hemlock*.

However, this objection is mistaken. It fails to respect the exhaustivity constraint on variable selection, and thereby equivocates between different background modal contexts. It further fails to respect exclusivity, and thereby runs together what should be different variables. Rectifying this illuminates the implicit proportionality of common sense causal claims.

First, the objection ignores the fact that proportionality, in requiring exhaustive and exclusive variables, is relative to modal context. Take the hemlock example just outlined. Importantly, this example and corresponding claim are under-defined.[13] Translated into interventionist terms, all that this description provides is that there is some variable that takes a value that represents Socrates drinking hemlock, and an intervention on this variable changes the value of some other variable to one that represents Socrates dying. But, a number of different variables could represent the purported cause, and a number of different models could represent its relationship to the effect of Socrates' dying. Which of these is accurate depends on what the relevant alternatives to drinking hemlock are. How these details get filled in will determine whether or not the variable that represents Socrates drinking hemlock is proportional.

I hold that the common sense claim that drinking hemlock causes Socrates's death implicitly takes the relevant alternative to be Socrates's *not* drinking hemlock. The default context is taken to be that hemlock was the only possible poison, and drinking it the only possible means of consumption. Given this context, the exhaustive variable would take the values {*drinks-hemlock*, *doesn't-*

[13] I take this to be common knowledge. See Franklin-Hall 2016; McDonnell 2017; and Weslake 2017

drink-hemlock}. But, such a variable is indeed proportional to the effect variable. Thus, the common sense cause is, in fact, proportional.

Such a defense requires that common sense claims be implicitly relative to a modal context. I'm not the first to relativize common sense claims to context. Philosophers such as Mackie and Schaffer make such a move, albeit with different ends in mind.[14] However, both McDonnell and Weslake explicitly deny this kind of relativity.[15] They claim that the very fact that we have strong and convergent intuitions about common sense examples, despite their being under-determined, demonstrates that the intuitions are not sensitive to filling in details.

In response, I argue that we respond to common sense causal examples in the same way that we respond to standard conversations. According to Grice, communication is governed by a set of conversational maxims.[16, 17] The maxims most relevant to how an audience engages with these under-defined causal examples are the maxims of *quantity* and *relation*. Taken together, these maxims enjoin an interlocutor to,

> Make your contribution as informative as is required (for the current purposes of exchange)….[and no] more informative than is required,….[and b]e relevant. (1989, 26 – 27)

Thus, the conversationally natural way to fill in the modal context of these examples is to take each fact as informative and relevant, and to assume that all informative facts have been provided.

The only information provided by the hemlock example is the following: (i) Socrates drinks hemlock. (ii) Socrates dies. The Gricean maxims tell us that this is all the information needed, and that nothing significant has been left out. So, the details are filled in as continuous with everyday life. In possible world speak,

[14] See Mackie 1974, especially chapter 2; and Schaffer 2005
[15] McDonnell 2017; Weslake 2017
[16] See Grice 1989
[17] Bontly makes a similar point (see 2005)

we're looking only at worlds which have a similar environment, a biologically similar Socrates, etc., and in which laws of metaphysical necessity hold.

The causal focus is on Socrates's drinking hemlock. This means that in evaluating the causal relationship, everything else is held fixed and the fact of the drinking hemlock is varied. Due to the absence of any other details, the only real alternative to Socrates's drinking hemlock is his not drinking hemlock. Nothing suggests that there are alternative means of consuming the hemlock. Further, it's not a common occurrence in everyday life to have alternative means of consuming a given poison. Treating *eating hemlock* as a relevant alternative would be to arbitrarily introduce something that wasn't otherwise specified, and whose presence can't be justified by everyday experience.

The objection from common sense assumes different possible alternatives than what I take to be implicit, and then tries to say that relative to these other alternatives, the common sense causal claim is not proportional. I have argued that the common sense cause is simply not relative to these other alternatives.

However, even given other possible alternatives, the common sense cause would still be proportional. The second mistake that the objection makes is that it fails to appreciate the constraint of exclusivity.

The objection holds that there is some relevant alternative to Socrates's drinking hemlock that preserves his consuming it. Take as an arbitrary alternative his eating hemlock. Socrates could both drink and eat the hemlock – he could wash down a hemlock salad with a glass of hemlock milk, for example. Following exclusivity, then, these possibilities should be represented by distinct variables – one that can take the value *drinks-hemlock*, call this *D*, and one that can take *eats-hemlock*, call this *E*.

But, now there is no problem. Following Woodward's response to early pre-emption cases,[18] we can hold *E* fixed at the value that represents Socrates not eating the hemlock, and see if the changes in *D* – which we can ensure meets exhaustivity by giving it the second value *doesn't-drink-hemlock* – line up with the changes in the effect variable. They do. When an intervention sets the value of the cause variable to *drinks-hemlock*, the effect variable takes the value *dies*. When an intervention sets the value of the cause variable instead to *doesn't-drink-hemlock*, the effect variable changes value to *doesn't-die*. Once again, the common sense cause is proportional.

If, on the other hand, the situation is such that Socrates's drinking hemlock is indeed mutually exclusive with his eating hemlock, then *drinks-hemlock* and *eats-hemlock* could be values of the same variable. Imagine that Socrates's jailor only has enough money to purchase either hemlock leaves or hemlock milk, but not both. In this case, neither Socrates's drinking nor his eating will be proportional. The proportional cause is instead his consuming hemlock. The proportional variable will therefore be one that takes as values {*consumes-hemlock*, *doesn't-consume-hemlock*}.

But, this is not in conflict with common sense – so long as we abstract away from normal everyday circumstances, and instead genuinely fix the situation as one in which Socrates is forced to consume hemlock, arbitrarily receiving hemlock leaves or milk. When, given this background, we're asked what causes Socrates's death, it is natural to say that it was his consuming hemlock. After all, it isn't the drinking nor the eating that makes a difference to whether Socrates dies, since had he not done one he would have done the other. It is his consuming hemlock rather than not.

Finally, I'd like to point out that the intuition that Socrates's consuming hemlock is the more proportional cause is actually misguided. The naïve intuition holds that an exhaustive and exclusive variable with the value *consumes-hemlock* – call this H_1 – is more proportional to the exhaustive and exclusive variable with the

[18] See Woodward 2003

value *drinks-hemlock* – call this H_2. But, the modal context to which H_1 will be exhaustive is different than that to which H_2 will be. They're therefore not even in competition for proportionality. Instead, I suggest that this intuition is a response to the fact that H_1's modal context is more inclusive than that of H_2. H_1 can accurately (and proportionally) represent the cause of Socrates's death in a wider range of situations than can H_2. But, this is about stability – as earlier defined – not about proportionality. The model that employs H_1 is simply *more stable* than that which employs H_2. This putative proportionality intuition is actually responding to the property of stability.

7. Conclusion

In this paper, I have defended the interventionist formulation of proportionality by explicating the exhaustivity and exclusivity constraints, and stipulating that proportionality requires variables that meet these constraints.

These constraints have been defined on the assumption that a variable represents a particular object's instantiations of a particular type of property. But, they are easily generalized to cover alternate objects of representation. Take events, for example. If variables represent particular kinds of events occurring or failing to occur, then exhaustivity would require that the values of a variable cover the entire range of possibilities of event occurrence for whatever type of event the variable represents. Exclusivity would require that the values of a variable be event occurrences such that no two could occur simultaneously.

Finally, I have articulated how the interventionist formulation of proportionality responds to the objection from common sense. Such an objection dissolves once the explicated constraints on variable selection are honored.

8. References

Bontly, Thomas. 2005. "Proportionality, Causation, and Exclusion." *Philosophia* 32 (1-4): 331 – 48

Franklin-Hall, Laura. 2016. "High-Level Explanation and the Interventionist's 'Variables Problem.'" *British Journal for the Philosophy of Science* 67 (2):553 – 77

Greenland, Sander, and Babette Brumback. 2002. "An Overview of Relations Among Causal Modelling Methods." *International Journal of Epidemiology* 31:1030 – 37

Grice, H. Paul. 1989. *Studies in the Way of Words.* Cambridge, MA: Harvard University Press

Hitchcock, Christopher. 1996. "The Role of Contrast in Causal and Explanatory Claims." *Synthese* 107 (3): 395 – 419

--- 2009. "Causal Modelling." In *The Oxford Handbook of Causation*, ed. Helen Beebee, Christopher Hitchcock, and Peter Menzies, 299 – 314. Oxford: Oxford University Press

Lewis, David. 1983 "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343 – 77

List, Christian, and Peter Menzies. 2009. "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106 (9): 475 – 502

Mackie, J. L. 1974. *The Cement of the Universe.* Oxford: Oxford University Press

McDonnell, Neil. 2017. "Causal Exclusion and the Limits of Proportionality." *Philosophical Studies* 174 (6): 1459 – 74

Menzies, Peter. 1996. "Probabilistic Causation and the Pre-emption Problem." *Mind* 105 (417): 85 – 117

Menzies, Peter, and Christian List. 2010. "The Causal Autonomy of the Special Sciences." In *Emergence in Mind*, ed. Cynthia Macdonald and Graham Macdonald, 108 – 28. Oxford: Oxford University Press

Papineau, David. 2013. "Causation is Macroscopic but not Irreducible." In *Mental Causation and Ontology*, ed. Sophie C. Gibb and Rögnvaldur Ingthorsson, 126 – 52. Oxford: Oxford University Press

Paul, L.A. 2000. "Aspect Causation." *The Journal of Philosophy* 97 (4): 235 – 56

Schaffer, Jonathan. 2005. "Contrastive Causation." *The Philosophical Review* 114 (3): 297 – 328

Shapiro, Larry, and Elliott Sober. 2012. "Against Proportionality." *Analysis* 72 (1): 89 – 93

Weslake, Bradley. 2013. "Proportionality, Contrast, and Explanation." *Australasian Journal of Philosophy* 91 (4): 785 – 97

--- 2017. "Difference-Making, Closure, and Exclusion." In *Making a Difference,* ed. Helen Beebee, Christopher Hitchcock, and Huw Price, 215 – 32. New York: Oxford University Press

Woodward, James. 2003. *Making Things Happen.* Oxford: Oxford University Press

--- 2008a. "Mental Causation and Neural Mechanisms." In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, ed. Jakob Hohwy & Jesper Kallestrup, 218 – 62. Oxford: Oxford University Press

--- 2008b. "Response to Strevens." *Philosophy and Phenomenological Research* 78 (1): 193 – 212

--- 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biological Philosophy* 25 (3): 287 – 318

--- 2015. "Interventionism and Causal Exclusion." *Philosophy and Phenomenological Research* 91 (2): 303 – 47

--- 2016. "The Problem of Variable Choice" *Synthese* 193 (4): 1047 – 72

Yablo, Stephen. 1992. "Mental Causation" The Philosophical Review 101 (2): 245 – 80

Species as Models

Abstract: This paper argues that biological species should be construed as abstract models, rather than biological or even tangible entities. Various (phenetic, cladistic, biological etc.) species concepts are defined as set-theoretic models of formal theories, and their logical connections are illustrated. In this view organisms relate to a species not as instantiations, members, or mereological parts, but rather as phenomena to be represented by the model/species. This sheds new light on the long-standing problems of species and suggests their connection to broader philosophical topics such as model selection, scientific representation, and scientific realism.

1 Introduction

Biological species has arguably been one of the most controversial topics in the philosophy of biology. Philosophers and biologists alike have long debated over "correct" concepts of species and their ontological status. The traditional account took species as a category, class, or type instantiated by individual organisms. After the advent of evolutionary theory, the typological concept came under fire by those who identify species with a part of biological lineage (Ghiselin 1974; Hull 1976). They forcefully

argued that a species is not an abstract type but a concrete historical entity of which individual organisms are mereological bits. Although this individualist thesis became a de-facto standard in the philosophy of biology in the last century, some have complained its lack of explanatory power and called for a revival of a type or natural-kind based concept of biological species (Boyd 1999).

To this debate between individualists and typologists, this paper introduces yet another thesis according to which species taxa are models of scientific theory. Model is a notoriously equivocal concept, but in this paper it is understood as a set-theoretic entity that makes sentences of a given theory true or false. This implies that biological species are mathematical, rather than biological or even tangible, entities. To work out this claim I begin Section 2 with a reconstruction of various (e.g., phenetic, cladistic, biological etc.) species concepts in terms of formal models that licence characteristic sets of inferences. The model-theoretic rendering illustrates logical connections among different species concepts and provides a platform to evaluate them as a problem of *model selection*. Section 3 then expounds on philosophical implications of the model-theoretic interpretation. Identifying species with models entails that the organism-species relationship is not instantial or mereological, but rather representational; i.e., species as models *represent* individual organisms. This opens the possibility of applying general philosophical discussions on scientific representation and realism to vexed questions concerning the epistemic and ontological status of biological species. Through these arguments this paper puts the species problem under broader contexts of model selection, scientific representation, and scientific realism, depicting it as a special case of the generic question as to how science investigates the world.

2 Species as models

This section fleshes out the main claim of this paper by reconstructing various species concepts as set-theoretic models. The central idea is that species concepts specify theories that underpin biological inferences and descriptions, and species are models that satisfy such theories.

2.1 Typological species concepts

The traditional typological view defines species by its essence, or necessary and sufficient conditions or traits. This finds a straightforward expression as a biconditional form $\forall x(Sx \leftrightarrow T_1x \wedge T_2x \wedge \cdots)$. The extension of species S that satisfies this formula then is the intersection $\bigcap_i \mathbf{T}_i$ (see Figure 1(a)).

Though crude as it is, the biconditional formulation allows certain inferences from traits to species and vice versa. It is this kind of logical reasoning that has enabled, for example, the famous French zoologist George Cuvier to reconstruct the anatomy of a whole organism from just a single piece of bone. As is well known, however, such inferences have very restricted validity, because in most cases it is impossible to find a definite set of phenotypic or genetic characteristics that exclusively defines a given species. Evolution implies species boundaries to be necessarily "fuzzy," which undermines simple biconditional forms. The typological species concept has thus been criticized for its lack of expression ability: a simple algebra of trait-sets cannot capture the nuanced reality of biological species.

(a) Typological Sets (b) Cluster Phenotypic space (c) Cladistic Posets (d) Relational Graphs

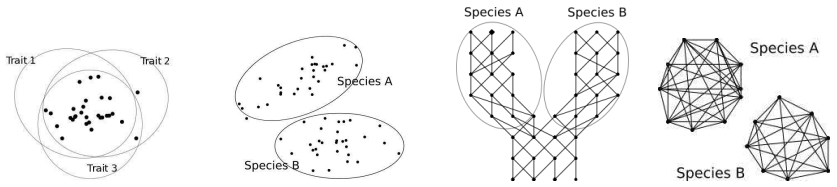


Figure 1: Illustrations of models of various species concepts, with corresponding formal setups. In each model dots/nodes represent individuals. See text for explanation.

2.2 Cluster species concepts

The cluster species concepts avoid this difficulty by defining a species as a group or cluster of similar organisms that do not necessarily share a common set of traits. The question then is how to define similarity. Its earliest variant, the phenetic species concept, represents organisms in a multi-dimensional space each axis of which defines a recorded trait (Sokal and Sneath 1963). Phenotypic similarity is then measured by the euclidean distance between two points/organisms, and a chunk or cluster of organisms in this euclidean space is identified as a species (Figure 1(b)). The choice of euclidean distance is not obligatory. One could, for example, measure similarity by the cosine between two points in the normalized phenotypic space, in which case the similarity amounts to correlation, with a species being identified as a correlated cluster or more generally a *probability distribution* over the phenotypic space (Boyd 1999).

The phenotypic space with a certain metric or probability distribution is certainly a much richer machinery than overlapping sets and allows for more nuanced expressions and inferences. The sophisticated theoretical background (euclidean geometry or

probability theory) enables one to measure the similarity among organisms and to make a trait-species inference in the absence of necessarily or sufficient criteria. To what extent such clustering and inference reflect objective species boundaries, however, was disputed, for the similarity calculation depends much on which phenotypic characters are taken into account. It should also be noted that, like the typological concept, the cluster concepts are purely static and lack a means to express the evolutionary past, the point often criticized by more historical approaches to species.

2.3 Cladistic species concepts

The cladistic species concepts focus on evolutionary history and define species solely in terms of phylogenetic relationships, as a "branch" (monophyletic group) in the evolutionary tree (Hennig 1966). Since ancestral relationship is antisymmetric and transitive, phylogeny forms a (strict) *partially ordered set* or *poset* $(\Omega, \prec)$, with Ω corresponding to a set of organism and $\prec$ meaning "is an ancestor of." A cladistic species is then defined as descendants from some founder organism(s) ω_f:

$$\{\omega \in \Omega : \omega_f \prec \omega\}. \tag{1}$$

An obvious advantage of the cladistic concepts is that it is faithful to the fact of evolution, and for this reason it has been most well received by biologists and philosophers alike. It is not, however, without flaws. For one, although the requirement of monophyly specifies a necessarily condition, it is silent as to how big a branch must be to qualify as a species (for even a small family can satisfy (1)), and so far no satisfactory sufficient condition was given (Velasco 2008). The monophyly requirement has also been

criticized to be too strong, for it would count birds as reptiles because the smallest monophylic group including lizards, snakes, and crocodiles also includes birds. That is, the cladistic species concepts make paraphyletic groups like reptilia *meaningless* (*Sensu* Narens 2007), which strikes some to be too high a price to pay.

2.4 Relational species concepts

Another popular approach is to define a species as a group of individuals in a certain relationship to each other. The biological species concept, for instance, defines species as "groups of interbreeding populations that are reproductively isolated form other such groups (Mayr 1942)" so that the required relationship here is mutual crossability. Other variants focus on reproductive competition (Ghiselin 1974) or organisms' capacity to recognize each other as a possible mate (Paterson 1985). All these proposals try to reduce species into mutual relationships (interbreeding, competition, recognition, etc.) between a pair of organisms. If we represent such relationships by an edge between nodes/organisms, a relational species can be defined as an isolated complete subgraph or *clique* in an undirected graph, that is, a group of nodes in which every two distinct nodes are connected but none is connected to outside (Figure 1(d)). Relational species thus find their model in graph theory, where edges represent the relation in question.

A common criticism of relational species concepts is that the focal relationship such as crossability sometimes fails to induce isolated cliques because some organisms at a species boundary can often mate with organisms that are thought to belong another species (e.g. ring species). Moreover, the biological species concept has been criticized to imply every asexually reproducing organism forms a distinct species (for any singleton

node is complete). These criticisms suggest that the real biological network is so "messy" that just a single relationship cannot divide it into distinct cliques in a non-trivial way.

2.5 "Combo" solutions

The model-theoretic rendering makes explicit what each species concept can and cannot meaningfully say about the biological world. Given that most of the criticisms we have seen concern the "cannot say" part, one way to deal with these difficulties is to combine different theories to obtain more complex definitions of species.

For instance, one may combine the cluster and cladistic species concepts and define a species as a *lineage that shares the same or similar phenotypic distribution*:

$$\{\omega \in \Omega : \omega_f \prec \omega \wedge \theta(\omega_f) = \theta(\omega)\} \tag{2}$$

where $\theta : \Omega \to \mathbb{R}^n$ assigns distribution parameters to each organism $\omega \in \Omega$.[1] On this definition one may meaningfully define paraphyletic species and distinguish birds from other reptiles on the basis of the difference in their phenotypic or genetic profiles. It can also account for anagenesis (speciation without branching) and continuity of species between a cladogenesis (splitting event).

If one replaces θ in (2) with a different function $\nu : \Omega \to N$ that maps organisms $\omega \in \Omega$ to their *niche* $\nu(\omega) \in N$, it becomes the *ecological species concept* which defines a species as "a lineage ... which occupies an adaptive zone minimally different from that of

[1]For non-parametric cases, we can set $\theta : \Omega \to \mathbb{R}^\infty$ and modify the definition as $\{\omega \in \Omega : \omega_f \prec \omega \wedge D(\theta(\omega_f), \theta(\omega)) < k\}$ where $D(\bullet)$ is a divergence measure (such as the Kullback-Leibler divergence) and k is a constant.

any other lineage in its range (Van Valen 1976, 233)."

Yet another combination is that of the cladistic and biological species concepts, which would define a species as a maximum monophylic lineage that can mutually interbreed, so that

$$\{\omega_x, \omega_y \in \Omega : \omega_f \prec \omega_x \wedge \omega_f \prec \omega_y \wedge \omega_x \sim \omega_y\} \tag{3}$$

where $\sim$ stands for crossability.[2] This will make up for the lack of a sufficient condition in the cladistic species concept, and accord well with the so-called *evolutionary species concept* which emphasizes the unique "evolutionary tendencies and historical fate" of each species (Wiley 1978, 17). It should be noted that this could also avoid the problem of ring species because two crossable organisms may not necessary share the same ancestor.

2.6 The scientific species problem as a problem of theory choice

The above discussion shows that (i) major species concepts can be defined as models of formal theories, and that (ii) more complex concepts can be obtained by combining basic ones. The model-theoretic approach characterizes each species concept with the formal apparatus it assumes, which in turn determines its expressive power or what can meaningfully be stated about organisms and/or their history (Narens 2007). In general, a richer theoretical apparatus allows for more nuanced expressions, which makes it less liable to counterexamples. This is illustrated in the progression from the typological to

[2]As in the case of the biological species concept, the crossability here must take into account the existence of two sexes.

cluster and then to cluster-cladistic concepts, where in each step the species concept acquires the ability to deal with fuzzy boundaries and evolutionary history, respectively.

It does not necessarily follow, however, that a richer concept is always desirable, because it tends to have a greater degree of freedom and requires more data in actual application. While only phylogenetic information suffices to demarcate cladistic species, the cluster-cladistic concept also requires phenotypic or ecological information, which in many cases may not be available. A stronger semantic power thus comes with a higher epistemic cost, as is often emphasized by pheneticists or cladists in their respective advocacy of the phenotypic cluster and cladistic species concepts.

This suggests that the competition among various species concepts should be understood as a problem of model selection, where different models are evaluated on the basis of their explanatory or descriptive power versus parsimony or operationality (Sober 2008). Indeed, most disputes among advocates of different species concepts arise from their differential emphasis on what aspects of the biological world a disable species concept needs and needs not take into account (Ereshefsky 2001), but the difficulty is that these emphases are often implicit and incommensurable. Although the model-theoretic approach does not arbitrate these debates, it provides a common formal framework that makes explicit the explanatory power and operationality of species concepts and facilitates evaluation of their respective advantage.

3 Philosophical implications

3.1 Species are models

Upon the model-theoretic reconstruction of various species concepts, we now turn to the philosophical thesis that species taxa should be construed as models proposed above, i.e., as set-theoretic entities. To proceed, let me first begin with an analogy from classical mechanics. Classical mechanics is a theory about Newtonian particles, which are customary defined as volumeless points or vectors in a three-dimensional Euclidean space. Newton's celebrated laws like $\mathbf{F} = m\mathbf{a}$ describe temporal evolution of a system composed of such "particles." This system is to be distinguished from any actual physical systems, say the solar system, for one thing, no concrete bodies are volumeless, nor do they indefinitely continue rectilinear motion as prescribed by Newton's first law. Newton's theory, or any other physical theories for that matter, is a description of idealized and abstracted models and not of actual phenomena (Cartwright 1983). That is, models of classical mechanics — which make its laws and statements true — are not concrete, physical entities, but rather abstract mathematical objects that can be constructed within set theory (McKinsey et al. 1953).

The role of models in science has been emphasized by the so-called semantic or model-based view of scientific theories (e.g. van Fraassen 1980; Suppe 1989).[3] In the traditional, logical-positivist view, a scientific theory was supposed to directly describe

[3]This label ("the semantic view") has been used to describe different, and logically independent, theses. In particular, while some philosophers (e.g. Suppes 2002) take a scientific theory as a *description* of models, others *identify* it with a set of models (van Fraassen 1980). In this paper I adopt the former thesis without committing to the latter.

observed data. This has set for positivists the difficult task of reducing theoretical concepts that seemingly lack direct empirical contents to observation vocabulary by way of *bridge laws* or *partial interpretations*. To avoid this difficulty, proponents of the model-based view take a model, rather than observation, as the primary descriptive target of a scientific theory. In this view, a theory specifies an abstract model that idealizes and extracts just salient factors, and only indirectly relates to actual phenomena via such an model.

I submit that the species problem is a variant of the positivist conundrum. Species is a highly theoretical concept, and various proposal of "species concepts" in the past can be understood as attempts to build bridge laws for reducing it to a set of observational or operational criteria. To date more than a dozen of different concepts have been proposed[4], with no general consensus — each has its own strength, but also weakness and exceptions when applied to the rich and heterogeneous biological world. The assumption has been that a species concept must be a faithful description of *actual* biological features or phenomena. But what if this assumption is untenable, or at least unreasonable? The model-based view has been quite popular among philosophers of biology (e.g. Beatty 1981; Lloyd 1988). If we adopt this view and construe evolutionary theory as describing models, then species too must be defined accordingly, i.e., as (a part of) abstract models that satisfy descriptions and/or inferences of the corresponding theory.

What, then, are theories about species? Without claiming to be exhaustive, this paper adopts Suppes's (2002) thesis that a scientific theory must be defined as a set-theoretical predicate. The foremost advantage of this approach is that it enables one

[4]Mayden (1997), for example, counts at least 22 concepts of species.

to easily harness a theory with mathematical apparatus necessary for sophisticated reasoning. As discussed above, contemporary studies on species rely heavily on quantitative methods to calculate similarity or reconstruct a phylogenetic tree from phenotypic or genetic data. Given that such mathematical reasoning requires matching formal models of calculus or probability theory, the straightforward way to define a species is to build it upon these mathematical backgrounds as an extension of these formal models. Section 2 is a preliminary sketch of applying this Suppesian program to various species concepts. If this attempt turns out to be successful, biological species are to be understood as parts of set-theoretic structures, just like Newtonian particles. That is, they are mathematical and abstract constructs, rather than physical or biological entities.[5]

The purpose of the set-theoretic exposition is not just to accommodate quantitative reasoning. Even with less quantitative cases like the biological species concept, it makes implicit assumptions explicit and suggests a way to deal with counterexamples. The problem of ring species, for example, arises from a conflict between the presumption that each biological species must be isolated and the fact that crossability is not necessarily transitive and thus fails to induce equivalence classes. One possible response to this charge then would be to weaken the former assumption and redefine a species just as a (not necessarily isolated) clique in the reproductive network. Clarification of theoretical assumptions helps us to assess other species concepts as well. For example, the phenetic species concept is often claimed to be "theory-free" in that it does not depend on any evolutionary hypothesis. But as we have seen in Sec. 2.2, the calculation of phenotypic

[5]Hence the present thesis should not be confused with the view that species are sets or collections of *organisms* (Kitcher 1984), which, after all, are concrete biological entities.

similarity presupposes a phenotypic space equipped with a particular (e.g., euclidean) metric, which is a fairly strong theoretical assumption. Also, cladists often stress the simplicity and purity of their monophylic species definition that only considers phylogenetic relationships. But in order to make use of likelihood methods to infer such relationships, as is common in practice, a simple poset is not enough: one also needs to assume some genetic or phenotypic distribution, and then there is no in-principle reason to exclude non-monophylic taxa from the definition of species (as (2) in Sec. 2.5).

The final but not least merit of the set-theoretic approach is its flexibility: it allows for a construction of a new species concept by combining existing ones (Sec. 2.5) or adding new theoretical assumptions. For instance, it is common in experimental biology to characterize a species by shared developmental or causal mechanisms: developmental biologists often talk about "the development of the chicken" and medical doctors rely on causal extrapolation when they prescribe a clinically-tested drug for their patient. Such a "causal species" may be defined by isomorphic *causal models*, which combine a probabilistic distribution and a causal graph over variables. Hence the discussion in Section 2 covers just a few samples that can be constructed within this general framework. This does not of course mean that every possible species concept can and must be formalized, but does suggest the potential of the set-theoretic approach to accommodate the use of existing species concepts and to develop novel ones.

3.2 Philosophical implications

Identifying species with theoretical models sheds new light on some vexed philosophical issues, one amongst which concerns how individual organisms are related to species taxa.

Philosophers have long debated whether the organism-species relationship is instantial (organisms are particular *instances* of a species *qua* class), membership (they are *members* of a species *qua* set; Kitcher 1984), or mereological (they are *parts* of a species *qua* genealogical entity; Ghiselin 1997). The model-theoretic approach suggests an alternative account, according to which a species *represents* (a group of) individual organisms. Just as the Rutherford-Bohr model represents the microscopic structure of atoms, models proposed in Section 2 represent biological populations: for example, nodes and edges consisting of the biological species model in Figure 1(d) respectively represent organisms and crossability. Representation captures our intuitive notion that a model and its target phenomenon share salient static or dynamic features up to a certain precision. Given that said, it must be admitted that the criteria and nature of scientific representation are diversified and still open questions (Frigg and Nguyen 2016). Hence calling the species-organism relationship representational does not necessarily demystify it, but at least implies that the problem is not endemic to evolutionary theory: it is rather a version of a broader philosophical issue as to how the use of scientific models help us understanding the world. This means that the arsenal of this rich philosophical literature can and should be consulted to elucidate the nature of the species-organism relationship. Another, more immediate implication is that the membership and mereological accounts must be both abandoned, for whatever the relationship between a model and phenomena turns out to be, the latter must certainly not be a member or part of the former.

Neither is representation identity or instantiation. Ideal gas is not identical to any actual gas, but only approximates thermodynamic characteristics of some. Hence strictly speaking it has no instantiation, but this does not detract its epistemic validity. Likewise

species concepts, as specifications of ideal models, need not directly apply to actual populations. No wild population big enough to qualify as a species would strictly satisfy the requirement of the biological species concept, because actual mating chance is often hindered by physiological, geographical, and other contingencies. In the same vein, a phenetic or genetic cluster is expected to have outliers when applied to a real population. However, the presence of such exceptions should not immediately invalidate the corresponding species concepts, because the value of a species concept consists less in its universal validity than its epistemic serviceability for inferences and explanations of evolutionary or biological phenomena. These two criteria often conflict: Cartwright (1983) even argues that explanatory theories necessarily distort the reality by idealizing the situation and extracting only relevant features, so that properly speaking they are "lies" by design. Cartwright's examples are physics and economics, but her idea also applies to the present context. The primary function of a species concept is to explain biological phenomena rather than to save them, so that a few discrepancies should not be taken as a falsification.

The conflict between exceptionlessness versus explanatory power also underlies the realism-nominalism debate over species. The proponents of the nominalistic thesis who claim a species to be nothing but a totality of individual organisms have motivated their view by criticizing the realist interpretation of species-as-class for its commitment to the typological thinking and failure to deal with the evident heterogeneity of biological phenomena (e.g. Ghiselin 1997). On the other hand, those who attach weight on the role of species concept in induction and explanation have upheld a realist position and treated species as natural kinds (Boyd 1999). The present thesis offers a third alternative, recognizing the explanatory role of species concept without committing to

the ontologically heavy assumption of natural kinds. As we have seen in Section 2, species as models licence particular sets of inferences. The cluster and typological species/models underpin an expectation that physiological or genetic features found in, say, laboratory animals would also be shared by other individuals of the same species, while the evolutionary species concept explains the reason of such intra-specific similarities. These explanations are effectuated by the same model representing numerically distinct individuals or phenomena to be explained. Note that this procedure no more presupposes the existence of the model as an independent, real entity, than do explanations based on, say, ideal gas. Indeed, explanations may be based on fictional models, as is the case with the Ising model in statistical mechanics.

This does not of course mean that models *must be* fictions, or that species do not exist. Recent advocates of scientific realism argue that successful scientific models capture some, especially structural, aspect of reality (Ladyman 2016). Given its affinity to the model-based view of scientific theories, species realists may well apply this line of reasoning to the present context, taking the set-theoretic structures as discussed in Section 2 as representing the reality or "essential feature" of biological species. Whether and to what extent such an argument carry over, however, remain to be examined by a further study.

4 Conclusion

The past debates over biological species have been based on the assumption that species concepts must describe actual biological phenomena, the strict adherence to which tends to rule out all but cladistic species as typological or inexact. The present paper

challenged this assumption and argued that the primary referent of a species concept is a (set-theoretic) model that licences a certain set of inferences specified by the concept. The model-theoretic rendering articulates explanatory power and theoretical assumptions of each species concept and illuminates logical relationships among them. Once species are specified as models, the long-standing competition among different species concepts reduces to a common problem of model selection. This suggests that evaluation of relative merits and demerits of species concepts must be based more on their explanatory power than on exceptionlessness.

On the philosophical side, the shift in the ontological status of species means that the organism-species relationship is not that of instantiation, membership, or mereology, but rather representation. The vexed issue that has troubled philosophers for decades, therefore, boils down to the broader problem as to how and why scientific models can be used to represent and explain the world. This suggests the possibility to apply the rich literature on scientific representation and realism to elucidate the epistemological and ontological nature of biological species.

In sum, the take home message of the present paper is that the species problem is not endemic to biology or evolutionary theory, but rather is a variant of general scientific and philosophical issues of model selection, scientific representation, and realism. The purpose of this paper was just to establish such a parallelism: determining its philosophical implications on specific debates such as realism or pluralism concerning biological species will be a task for future studies.

References

Beatty, John. 1981. "What's Wrong with the Received View of Evolutionary Theory?." *PSA 1980* 2: 397–426.

Boyd, Richard N. 1999. "Homeostasis, species, and higher taxa." In *Species: New Interdisciplinary Essays.* ed. Robert A Wilson, 141–158, Cambridge, MA: MIT Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Ereshefsky, Marc. 2001. *The Poverty of the Linnaean Hierarchy.* Cambridge: Cambridge University Press.

van Fraassen, Bas C. 1980. *The Scientific Image.* Oxford: Oxford University Press.

Frigg, Roman, and James Nguyen. 2016. "Scientific Representation." In *The Stanford Encyclopedia of Philosophy.* ed. Edward N Zalta, Metaphysics Research Lab, Stanford University.

Ghiselin, Michael T. 1974. "A Radical Solution to the Species Problem." *Society of Systematic Biologists* 23: 536–544.

——— 1997. *Metaphysics and the Origin of Species.* Albany, NY: State University of New York Press.

Hennig, Willi. 1966. *Phylogenetic Systematics.* Urbana, IL: University of Illinois Press.

Hull, David L. 1976. "Are species really individuals?" *Systematic Zoology* 25: 174–191.

Kitcher, Philip. 1984. "Species." *Philosophy of Science* 51: 308–333.

Ladyman, James. 2016. "Structural Realism." In *The Stanford Encyclopedia of Philosophy.* ed. Edward N Zalta, Metaphysics Research Lab, Stanford University.

Lloyd, Elisabeth A. 1988. *The Structure and Confirmation of Evolutionary Theory.* Princeton, NJ: Princeton University Press.

Mayden, R L. 1997. "A hierarchy of species concepts: the denouement in the saga of the species problem." In *Species The Units of Biodiversity.* ed. M F Claridge, H A Dawah, and M R Wilson, 381–424, London: Chapman & Hall.

Mayr, Ernst. 1942. *Systematics and origin of species.* New York, NY: Columbia University Press.

McKinsey, John C C, Patrick Suppes, and A C Sugar. 1953. "Axiomatic Foundations of Classical Particle Mechanics." *Journal of Rational Mechanics and Analysis* 2: 253–272.

Narens, Louis. 2007. *Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science.* Mahwah, NJ: Lawrence Erlbaum Associates.

Paterson, Hugh E H. 1985. "The Recognition Concept of Species." In *Species and Speciation.* ed. E. S. Vrba, 21–29, Pretoria.

Sober, Elliott. 2008. *Evidence and Evolution.* Cambridge: Cambridge University Press.

Sokal, Robert R, and Peter H A Sneath. 1963. *Principles of Numerical Taxonomy.* San Francisco, CA: W. H. Freeman and Co.

Suppe, Frederick. 1989. *The Semantic Conception of Theories and Scientific Realism.*: University of Illinois Press.

Suppes, Patrick. 2002. *Representation and Invariance of Scientific Structures.* Stanford, CA: CSLI Publication.

Van Valen, Leigh. 1976. "Ecological Species, Multispecies, and Oaks." *Taxon* 25: 233–239.

Velasco, Joel D. 2008. "The internodal species concept: a response to 'The tree, the network, and the species'." *Biological Journal of Linnean Society* 93: 865–869.

Wiley, Edward O. 1978. "The Evolutionary Species Concept Reconsidered." *Systematic Biology* 27: 17–26.

Historical Inductions Meet the Material Theory

by Elay Shech

Oct. 2018
(Pre-conference version)

Forthcoming in *Philosophy of Science*

Acknowledgements: I am indebted to John Norton and Moti Mizrahi for extremely valuable discussion and comments on earlier drafts of this paper. Thank you also to helpful conversation with the audience at the Auburn University Philosophical Society in the Spring of 2018 and participants in Gila Sher's *Truth and Scientific Change* reading group in the Fall of 2017 at the Sidney M. Edelstein Center for History and Philosophy of Science, Technology and Medicine at the Hebrew University of Jerusalem.

Abstract: Historical inductions, viz., the pessimistic meta-induction and the problem of unconceived alternatives, are critically analyzed via John D. Norton's material theory of induction and subsequently rejected as non-cogent arguments. It is suggested that the material theory is amenable to a local version of the pessimistic meta-induction, e.g., in the context of some medical studies.

1. Introduction

My goal is to contribute to a growing literature that is critical of historical inductions such as the pessimistic (meta-)induction (PMI) argument (Poincaré 1952, 160; Putnam 1978, 25; Laudan 1981) and the problem of unconceived alternatives (Stanford 2001, 2006) against scientific realism, concentrating mostly on the former. The PMI can be construed in different ways (Mizrahi 2015, Wray 2015), viz., as a deductive *reductio ad absurdum* (e.g., Psillos 1996, 1999), a counterexample to the no miracles argument and inference to best explanation argument for scientific realism (e.g., Saatsi 2005, Laudan 1981), or, usually, as an inductive argument (e.g., Poincaré 1952, Putnam 1978, Laudan 1981, Rescher 1987). In the following I will argue against the inductive version of PMI—or any construal of the PMI that makes use of historical induction—using John D. Norton's material theory of induction (Norton 2003, Manuscript). The upshot is that one ought to be critical of historical inductions that seem to fit the general form or pattern of a good inductive argument, but may in fact lack inductive warrant and force. Various critiques have been put against the PMI (e.g., Lange 2002, Lewis 2001, Mizrahi 2013), along with some defenses (e.g., Saatsi 2005). In Section 2 I will present the PMI and briefly discuss some criticism in order to place my own analysis in broader context. Section 3 presents the material theory of induction and argues that it dissolves the PMI, while Section 4 extends such claims to the more recent problem of unconceived alternatives. In Section 5 I note that the material theory of induction does leave room for a local version of the PMI, which holds in some

limited domain, such as in relation to certain medical studies (Ruhmkorff 2014). I end in Section 6 with a short conclusion.

2. The (Inductive) Pessimistic (Meta-)Induction

The modern formulation of the PMI is usually attributed to Laudan (1981) who argued that having genuinely referential theoretical and observational terms, or being approximately true, is neither necessary nor sufficient for a theory being explanatory and predictively successful. More generally, Anjan Chakravartty characterizes the argument as follows:

> [PMI can] be described as a two-step worry. First, there is an assertion to the effect that the history of science contains an impressive graveyard of theories that were previously believed [to be true], but subsequently judged to be false . . . Second, there is an induction on the basis of this assertion, whose conclusion is that current theories are likely future occupants of the same graveyard. (Chakravartty 2008, 152)[1]

The PMI then may take the following form:

> **[Inductive Generalization PMI]**
>
> P(i) Past theory 1 was successful but not genuinely referential or approximately true.
> P(ii) Past theory 2 was successful but not genuinely referential or approximately true.
> …
> C) Therefore, current (and perhaps future) theories are successful but (by induction) probably not genuinely referential or approximately true.

Laudan (1981) suggests that the history of science contains a graveyard of theories that were previously believed to be approximately true and genuinely referential, but that subsequently were judged to be false and not to refer. Estimations of the number of such superseded theories have been debated (e.g., Lewis 2001, Wray 2013) and recently Mizrahi (2016) presents evidence that challenges the "history of science as a graveyard of theories" claim. Others voice concerns regarding the period of history of science used in order to extract historical evidence (e.g., Lange 2002, Fahrbach 2011) or the proper unit of analysis, i.e., theories vs. theoretical entity (e.g., Lange 2002, Magnus and Callender 2004). Similarly, Park (2011, 83) and Mizrahi (2013, 3220-3222) have argued that the PMI is fallacious due to cherry-picking data, biased statistics, and non-random sampling.

My own criticism of the inductive PMI comes from a different avenue. I will assume that the anti-realist does have randomly sampled historical evidence from the correct period of history and with the proper unit of analysis (whatever those

[1] cf. Wray (2015, 61).

may be) that is not biased or cherry-picked. Still, on the material theory of induction the PMI will not be a cogent argument. In other words, I aim to identify what I take to be a more fundamental (although not categorically different) problem with the PMI.

3. PMI Meets the Material Theory

3.1 The Material Theory of Induction in a Nutshell

Consider the following formally identical inductive inferences (Norton 2003, 649):

> P1) Some samples of the element bismuth melt at 271 degrees C.
> C1) Therefore, all samples of the element bismuth melt at 271 degrees C.
>
> P2) Some samples of wax melt at 91 degrees C.
> C2) Therefore, all samples of wax melt at 91 degrees C.

What makes the first argument an inductively strong and cogent argument while the second a weak and non-cogent inductive argument? Norton (2003, Manuscript) has argued that formal theories of induction, which provide universal schemas that are meant to identify the inductions that are licit and those that are not, stand against an insurmountable difficulty when facing such a question.[2] Instead, he offers a material account of induction:

> In a material theory, the admissibility of an induction is ultimately traced back to a matter of fact, not to a universal schema. We are licensed to infer from the melting point of some samples of an element to the melting point of all samples by a fact about elements: their samples are generally uniform in their physical properties. … *All inductions ultimately derive their licenses from facts pertinent to the matter of the induction*. (Norton 2003, 650; original emphasis)

Norton calls the local facts that power inductive inferences "material postulates." Material postulates themselves are supported by other instances of induction that are licensed by different material postulates.

3.2 Material Analysis of PMI

Many of the criticism of the inductive PMI discussed above amount to the claim that the universal schema used by the likes of Laudan (1981), namely, (P3) Some A's are B's, (C3) Therefore, all A's are B's, does not apply in the case of the PMI because various criteria needed to implement the scheme, e.g., random sampling, correct historical period, proper unit of analysis, have not been met. What I wish to do here

[2] I will not defend Norton's theory or claims here. He dedicates an entire book to the matter in Norton (Manuscript).

is conduct a material analysis of the PMI. Considering the above presentation of the PMI in its [Inductive Generalization PMI] form we may ask, what powers the inductive inference, i.e., what material postulate licenses the pessimistic conclusion?

In context of the two inductive arguments considered in Section 3.1, we note that there is no material postulate that licenses the inductive inference in the case of wax (P2 too C2) but there is one in the case of bismuth (P1 to C1): Generally, chemical elements are uniform in their physical properties. By analogy, the presumption of the meta-induction is that each historical case study looked at is an instance of the same thing, a discovery of induction in science. If we are to perform the meta-induction then there needs to be something in the background facts that unifies all such inductions, just like the fact chemical elements are generally uniform in their physical properties warrants the inductive inference regarding the melting point of bismuth. Let us consider several options.

First, perhaps the material fact is that most scientists use a common rule or method in constructing or discovering successful theories, something along the lines of Mill's methods of experimental inquiry in his *System of Logic* (1872, Book III, Ch. 7). If so, the properties of the rule would be used to authorize the induction. Is there such a rule, or perhaps, some common scientific method? A glance at the history of science suggests that this is unlikely. Newton's deduction from the phenomena, is very different from Darwin's inference to best explanation, which in turn differs radically from Einstein's thought experiments with lights beams, trains, and elevators.[3] More generally, there seems to be a consensus among historians and philosophers of science that something like "the scientific method" is really more of an umbrella term for very different methods used by scientists to construct and discover theories. After all, novel problems necessitate novels solutions, and the commonality that does arise in different cases, say, attempts to minimize error or to be objective, is not the kind of commonality that we seek in powering the PMI and drawing the pessimistic conclusion. For instance, in his book *Styles of Knowing: A New History of Science from Ancient Times to the Present*, Chungling Kwa (2011) argues that there is no single, fundamental method used in science: "there is not just one form of Western scientific rationality; there are at least six." The framework of six "styles of knowing," includes the deductive, the experimental, the hypothetical-analogical, the taxonomic, the statistical, and the evolutionary style, and is based on Alistair Crombie's (1994) three-volume work *Styles of Scientific Thinking*. Similar, Ian Hacking (also taking lead from Crombie's work) has argued that there are distinct "styles of reasoning" used in science, such as the postulational style, the style of experimental exploration, the style of hypothetical construction of models by analogy, the taxonomic style, the statistical style, the historical derivation of genetic development, and the laboratory style (Hacking 1992). This further

[3] In fact, see Norton (Manuscript, Ch. 8-9) who argues that even in historical cases where the *same* principle is applied by scientists, viz., inference to best explanation, "at best we can find loose similarities that the canonical examples of inference to best explanation share," so that no common rule of the kind needed to power the PMI can be found (Ch. 8, p. 1).

corroborates the idea that scientific methods used for theory construction and discovery, as well as for scientific explanation, are very diverse.

More generally, scientific theories are not kind of things that portray the type of uniformity needed to license inductive inferences on Norton's material theory. Albeit in a different context, a similar point is nicely made by Mizrahi (2013, 3218):

> A uniform—as opposed to diverse—sample might be a sample of, say, copper rods. From a sample of just a few copper rods that are tested for electrical conductivity, it is reasonable to conclude that all copper rods conduct electricity because, if you have seen one or two copper rods, you have seen them all (given their uniform atomic structure). Scientific theories, however, are not as uniform as copper rods. The point, then, is that any sample of theories is not going to be uniform in a way that is required for a "seen one, seen them all" inductive generalization.

Similarly, and second, perhaps there are some facts about investigating scientist themselves, how they work, and/or the problems situations that they work in, which can unify the historical evidence in a way that provides us with the inductive warrant we seek. Maybe such facts will include something about the psychology of scientists: their fastidiousness and fear of error, their facility at jumping to conclusions, or perhaps their curiosity, logic, creativity, skepticism, etc. However, in a similar manner to the search for a common rule used in constructing successful theories, the history of science furnishes us with scientists that are heterogeneous enough in their psychological traits, and work in such varied contexts, so as not to provide us with any was to unify the various historical cases in a way pertinent to licensing the pessimistic inference of the PMI.

Third, perhaps we can circumvent looking to a common rule of constructing or discovering theories, or searching for common traits among scientists, by noting that the follow candidate material postulate would power the PMI:

> MP-PMI: Generally, successful theories are not genuinely referential and/or approximately true.

But how would we establish MP-PMI? One option is to appeal to the PMI itself, but this would either be circular or else push us to look for another material postulate. Another option is just to grant the MP-PMI as a reasonable assumption. Perhaps anti-realists or instrumentalists would think that this is a sensible starting point, but their target realist opponent would surely reject such an assumption as question begging. Last, perchance there is some fact about explanatory and/or predictively successful theories that renders them, generally, not genuinely referential and/or approximately true? Possibly part of the essence of successful theories is to misrepresent the world? To me this seems highly unlikely and at odds with any levelheaded intuition but, in any case, if we could argue that successful theories are essentially inaccurate then we would not need the PMI in the first place!

Fourth, we may want to construe the PMI in its inductive generalization form as a kind of abductive argument with the following type of material postulate:[4]

> **[Inductive Generalization PMI – Abductive version]**
>
> P(i): The success of past theory 1 (constructed using method m) is not best explained by its truth.
> P(ii): The success of past theory 2 (constructed using method m) is not best explained by its truth.
> …
> MP: Scientific theories constructed using method m are generally uniform with respect to what best explains their predictive success.
> C: The success of our best current (and perhaps futures) theories (constructed using method m) are not best explained by their truth.

Stating the PMI as above has the merit of directly engaging with the "no miracles argument" for scientific realism, namely:

> That terms in mature scientific theories typically refer [to things in the world] …, that theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when it occurs in different theories—these statements are viewed by the scientific realist not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate scientific description of science and its relations to its objects. (Putnam 1975, 73)

But worries abound. First, the realist may very well deny P(i), P(ii), etc., and argue that the success of past theories is best explained by their truth but that, as it turns out, either the best explanation did not hold in this case or else there is some sense in which past theories, insofar as they were successful, were approximately true or on the road to truth. Second, construing the argument as an abduction opens up a Pandora's box of problems associated with the notion of explanation: What is explanation? Are there accounts of explanation where success is best explained by truth and ones in which it isn't and, if so, which account of explanation is relevant in this context? And so on.

Third, the cogency of the argument depends on the idea that all theories appealed to were constructed with some method m, but we already judged that there is no one method that is relevant to constructing scientific theories. Perhaps phenomenological models are good candidates for the type of things that can provide empirical success but are not generally approximately true.[5] Thus, at best, the above argument can power a kind of local PMI: Successful theories constructed

[4] Thanks to Tim Sundell for suggest this line of thought.
[5] Phenomenological models are, generally, not considered explanatory.

by method m are not approximately true. We'll consider one such case in more detail in Section 5.

In short, on the material theory of induction inductive arguments are powered by facts, by material postulates, but in the context of the PMI it seems unlikely that any such non-question begging postulates, which wouldn't render the PMI obsolete, can be found. This is so even if, say, the historical data was not cherry-picked, and the right unit of analysis and correct period of history were used. In other words, I'm equally skeptic of projects that attempt to block the pessimistic conclusion by, for example, taking a random sample of past scientific theories, e.g., Mizrahi (2016). In the following section I'll attempt to extend such claims to the problem of unconceived alternatives.

4. Extension to the Problem of Unconceived Alternatives

Recently, P. Kyle Stanford (2001, 2006) has developed what may be characterized as a new version of the PMI:

> … I propose the following New Induction over the History of Science: that we have, throughout the history of scientific inquiry and in virtually every field, repeatedly occupied an epistemic position in which we could conceive of only one or a few theories that were well-confirmed by the available evidence, while subsequent history of inquiry has routinely (if not invariably) revealed further, radically distinct alternatives as well-confirmed by the previously available evidence as those we were inclined to accept on the strength of that evidence. (Stanford 2001, S8-S9)

The problem of unconceinved alternatives as an argument against scientific realism has been criticized on various grounds (e.g., Chakravartty 2008, Devitt 2011, Mizrahi 2015), but my goal here is just to note that the discussion of Section 3 can be extended to this new version of the PMI, which can be construed as follows:

> P(i) In the past time of theory 1, theory 1 was successful but there were unconceived alternative theories that were as well supported by available evidence but with radically different ontology.
> P(ii) In the past time of theory 2, theory 2 was successful but there were unconceived alternative theories that were as well supported by available evidence but with radically different ontology.
> …
> C) Therefore, in present times, current theories are successful but (by induction) there probably are unconceived alternative theories that are as well supported by available evidence but with radically different ontology.

What we need for the material analysis is something like: Generally, successful theories are underdetermined by data due to possible unconceived alternative theories. In a similar fashion to the MP-PMI, we could look to some common rule used by scientists to conceive theories, or some common psychological traits among

scientist, that may ground the idea that successful theories are such that empirically adequate unconceived alternatives always exists. But for the same reasons discussed above, it seems unlikely that any such common rule or traits will be found. That said, perhaps cognitive facts about human scientists might support the inductive inference to the conclusion that we always miss some alternative theories, which in turn are consistent with the available evidence. What is attractive about this line of thought is that it does seem plausible that due to our cognitive limitations there are always "unconceived alternatives." However, mere cognitive limitations do not support the further conclusion that there are unconcieved alternative theories that are *consistent with available evidence*.

Alternatively, one may think that Stanford's new induction circumvents the material objection: modal reflections alone convince us that there are always unconceived alternative theories that can explain and predict empirical phenomena just as well or better than conceived theories. But how can we come to such a conclusion based on modal reflections alone? Isn't it conceivable if not possible that there would be a point in history with no unconceived alternatives and isn't conceivable if not possible that we are at such point in time in history? Moreover, it is unclear what to make of theory-independent modal claims (unless one has logical modality in mind, which isn't the case here). Certainly, we can talk about different physically possible worlds given a particular physical theory. For instance, various solutions to the Einstein field equations are taken to denote different possible universes according to relativity theory. But it isn't clear what is meant by different possible or alternative conceivable *theories* given no meta-theory as a constraint, so to speak.[6] In any case, if we know that unconceived alternative theories always exist based on modal reflections alone, then the historical induction is doing no work for us at all.

5. Room for a local, material pessimistic induction?

Although the material analysis given here may prompt us to be skeptical of historical inductions (insofar as one is moved by the material theory of induction), it can help us understand why *local* pessimistic inductions may be tenable. Specifically, I want to look at a recent discussion by Rumkorf (2014) who contends that meta-analyses in medicine such as Ioannidis' (2005a, 2005b), which show that a disconcertingly high percentage of prominent medical research findings are refuted by subsequent research, can be developed into a local pessimistic induction. Ioannidis (2005a, 2005b) is concerned with studies, denoted "M-studies," that satisfy the following criteria: "being highly cited, using contemporary research and statistical methods, and being among the first studies to investigate a question at issue" (Rumkorf 2014, 420). Rumkorf's (2014, 421) then uses the various conclusions of Ioannidis (2005a, 2005b) to generate a local PMI in the field of medicine (PMI-M):

[6] What would count as a (logically possible but physically) impossible theory in such a context?

> E1 41% of the associative or causal claims made by M-studies in the sample were inconsistent with the results of subsequent published studies either (1) because the later studies provided evidence against the existence of the association or effect; or (2) because the later studies provided evidence that the magnitude of the association or effect was significantly different.
>
> E2 Therefore, we can expect approximately 41% of the associative and causal claims made by M-studies to be inconsistent with subsequent published studies.

On Norton's theory we need to appeal to a material postulate to license the pessimistic inductive inference in the transitions from E1 to E2, but since we are now working in a limited domain without many heterogeneous examples as in the whole history of science, we may now find some significant commonality between the methods used in different M-studies that can act as licensing facts. What are the background facts that power the PMI-M? Here are some options extracted from Ioannidis's diagnosis of his meta-analysis and quoted in Ruhmkorf (2014, 219):

> Contributing factors include: bias in research (Ioannidis 2005b); non-randomized trials (Ioannidis 2005a); smaller rather than larger sample sizes in refuted studies (Ioannidis 2005a, 224); and publication and time-lag biases (whereby studies with highly significant and potentially aberrational positive results are overrepresented among published articles in major journals and are published more quickly than other articles) (Ioannidis 2005a, 224). Particularly intriguing is the idea that large-scale features of the structure of medical and biological inquiry contribute to the high contradiction rate. Having a number of distinct working groups looking at the same problem increases the chances that at least one of them will find something statistically significant, especially if they are looking at a wide array of possible relationships (Ioannidis 2005b, 697–698). The computational power and richness of data sets available to researchers increases the chance that some of them will be successful in achieving statistical significance, even when no real relationship exists (Ioannidis 2005b, 701).[7]

These various factors, insofar as they are common to most M-studies, are the type of background facts that warrant the pessimistic induction from a material point of view. One may worry of course that the pessimism associated with local PMI generalizes since, presumably, facts about biases and the like are facts about researchers in general, not just researchers in medical science in particular. But, although all scientific studies have to deal with challenges such bias, it may be the case that a particular local subfield, due to its specific nature and whatever social

[7] It should be noted that there are some problems with Ioannidis's (2005a, 2005b) methodology, as identified in Ruhmkorff (2014, 419-421), but they do not seem to be problematic enough to render the PMI-M not cogent.

norms are in place for collecting and disseminative evidence, is especially challenged in a way that can justify the pessimistic induction. The above suggests that this is indeed the case for M-studies.

To end, Ruhmkorff (2014) argues against global PMI on independent grounds (namely, he argues that the PMI commits a statistical error previously unmentioned in the literature and is self-undermining), and but he also argues for the plausibility of a local PMI, viz., M-PMI, and contends that there are clear advantages of PMI-M over PMI. What I wish to note here is that an additional advantage of PMI-M, or local pessimistic induction generally speaking, is that whereas global PMI dissolves upon a material analysis, a material account of PMI-M does seem viable.

6. Conclusion

I have argued that historical inductions such as the (global) PMI and the problem of unconceived alternatives dissolve if we work with the material theory of induction. The reason is that we lack the material postulates needed to license the pessimistic inference: the great heterogeneity of case studies from the history of science of conceiving, constructing, and discovering (explanatory and predictively successful) theories, along with abundant variety of context that scientists find themselves in and traits that they exhibit, make it unlikely that any commonality will be found strong enough to authorize the induction. One may of course object: so much worse for the material theory of induction! This is a fair point, but there is a more general moral to consider. In various situations one may be able to appeal to the notion of "induction" without much being at stake, but in the context of historical inductions like the PMI and problem of unconceived alternatives "induction" is doing a lot of (philosophically) heavy lifting and so the situation rightful calls for scrutiny. Such scrutiny has led to the various discussed criticism that are presented in the context of more traditional, non-material theories of induction. Accordingly, it seems appropriate to show that—even if we assume randomly sampled historical evidence from the correct period of history and with the proper unit of analysis that is not biased or cherry-picked, with no statistical error, etc.—historical inductions do not fare well on the material side of things. I leave objections to the effect that one ought to construe the PMI as a deductive argument, or through a different framework for induction, e.g., via hypothetical or probabilistic induction, for future work.

References

Chakravartty, A. 2008. "What You Don't Know Can't Hurt You: Realism and the Unconceived." *Philosophical Studies* 137: 149–158.

Crombie, A. C., 1995. *Styles of Scientific Thinking in the European Tradition*, 3 vols. London: Duckworth.

Devitt, M. 2011. "Are Unconceived Alternatives a Problem for Scientific Realism?" *Journal for General Philosophy of Science* 42: 285–293.

Fahrbach, L. 2011. "How the Growth of Science Ends Theory Change." *Synthese* 180: 139–155.

Hacking, I. 1992. "'Style' for historians and philosophers." *Studies in History and Philosophy of Science*, 23(1), 1–20.

Ioannidis, J. P. A. 2005a. "Contradicted and Clinically Stronger Effects in Highly Cited Clinical Research." *Journal of the American Medical Association* 294: 218–228.

Ioannidis, J. P. A. 2005b. "Why Most Published Research Findings Are False."*PLoS Medicine 2*: 696–701.

Kwa, C. 2011. *Styles of Knowing: A New History of Science from Ancient Times to the Present*. Pittsburgh: University of Pittsburgh Press.

Lange, M. 2002. "Baseball, Pessimistic Inductions, and the Turnover Fallacy." *Analysis* 62: 281–285.

Laudan, L. 1981. "A Confutation of Convergent Realism." *Philosophy of Science* 48: 19–49.

Lewis, P. J. 2001. "Why the Pessimistic Induction Is a Fallacy." *Synthese* 129: 371–380.

Magnus, P. D., and C. Callender. 2004. "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71: 320–338.

Mill, J. S. [1872] 1916. *A System of Logic: Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. 8th ed. London: Longman, Green, and Co.

Mizrahi, M. 2013. "The Pessimistic Induction: A Bad Argument Gone too Far." *Synthese* 190:3209–3226.

Mizrahi, M. 2015. "Historical Inductions: New Cherries, Same Old Cherry-picking." *International Studies in the Philosophy of Science* 29: 129–148.

Mizrahi, M. 2016. "The history of Science as a Graveyard of Theories: A Philosophers' Myth?" *International Studies in the Philosophy of Science* 30: 263-278.

Norton, J. D. 2003. "A Material Theory of Induction." Philosophy of Science 70: 647–670.

Norton, J. D. Manuscript. *The Material Theory of Induction*. See http://www.pitt.edu/~jdnorton/papers/material_theory/material.html

Park, S. 2011. "A Confutation of the Pessimistic Induction." *Journal for General Philosophy of Science* 42: 75–84.

Poincaré, H. [1902] 1952. *Science and Hypothesis*. New York: Dover. Originally published as *La science et l'hypothèse*. Paris: Flammarion.

Putnam, H. 1978. Meaning and the Moral Sciences. London: Routledge and Kegan Paul.

Psillos, S.: 1996, 'Scientific Realism and the 'Pessimistic Induction' ', *Philosophy of Science* 63 (Proceedings), S306–S314.

Psillos, S. 1999. Scientific Realism: How Science Tracks Truth. London: Routledge.

Rescher, N. 1987. Scientific Realism: A Critical Reappraisal. Dordrecht: D. Reidel.

Ruhmkorff, S. 2013. "Global and Local Pessimistic Meta-inductions." *International Studies in the Philosophy of Science* 27: 409–428.

Saatsi, J. 2005. "On the Pessimistic Induction and Two Fallacies." *Philosophy of Science* 72: 1088–1098.

Sklar, L. M. (2003). "Dappled theories in a uniform world." *Philosophy of Science*, 70, 424–441.

Stanford, P. K. 2001. "Refusing the Devil's Bargain: What Kind of Underdetermination Should We take Seriously?" *Philosophy of Science* 68 (Proceedings): S1–S12.

Stanford, P. K. 2006. Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford: Oxford University Press.

Wray, K. Brad. 2015. "Pessimistic Inductions: Four Varieties." *International Studies in the Philosophy of Science* 29: 61–73.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -510-

To be presented at the *2018 PSA Meeting*:

Can Quantum Thermodynamics Save Time?

Noel Swanson*

Abstract

The *thermal time hypothesis (TTH)* is a proposed solution to the problem of time: every statistical state determines a thermal dynamics according to which it is in equilibrium, and this dynamics is identified as the flow of physical time in generally covariant quantum theories. This paper raises a series of objections to the TTH as developed by Connes and Rovelli (1994). Two technical challenges concern the implementation of the TTH in the classical limit and the relationship between thermal time and proper time. Two more conceptual problems focus on interpreting the flow of time in non-equilibrium states and the lack of gauge invariance.

1 Introduction

In both classical and quantum theories defined on fixed background spacetimes, the physical flow of time is represented in much the same way. Time translations correspond to a continuous 1-parameter subgroup of spacetime symmetries, and the dynamics are implemented either as a parametrized flow on statespace (Schödinger picture) or a parametrized group of automorphisms of the algebra of observables (Heisenberg picture). In generally

*Department of Philosophy, University of Delaware, 24 Kent Way, Newark, DE 19716, USA, nswanson@udel.edu

covariant theories, where diffeomorphisms of the underlying spacetime manifold are treated as gauge symmetries, this picture breaks down. There is no longer a canonical time-translation subgroup at the global level, nor is there a gauge-invariant way to represent dynamics locally in terms of the Schrödinger or Heisenberg pictures. Without a preferred flow on the space of states representing time, the standard way to represent physical change via functions on this space taking on different values at different times, also fails. This is the infamous *problem of time*.

Connes and Rovelli (1994) propose a radical solution to the problem: the flow of time (not just its direction) has a thermodynamic origin. Equilibrium states are usually defined with respect to a background time flow (e.g., dynamical stability and passivity constraints reference a group of time translations). Conversely, given an equilibrium state one can derive the dynamics according to which it is in equilibrium. Rovelli (2011) exploits this converse connection, arguing that in a generally covariant theory, *any* statistical state defines a notion of time according to which it is an equilibrium state. The *thermal time hypothesis (TTH)* identifies this state-dependent thermal time with physical time. Drawing upon tools from Tomita-Takesaki modular theory, Connes and Rovelli demonstrate how the TTH can be rigorously implemented in generally covariant quantum theories.

The idea is an intriguing one that, to date, has received little attention from philosophers.[1] This paper represents a modest initial attempt to sally forth into rich philosophical territory. Its goal is to voice a number of technical and conceptual problems faced by the TTH and to highlight some tools that the view has at its disposal to respond.

2 The Thermal Time Hypothesis

We usually think of theories of mechanics as describing the evolution of states and observables through time. Rovelli (2011) advocates replacing this picture with a more general *timeless* one that conceives of mechanics as describing relative correlations between physical quantities divided into two classes, *partial* and *full* observables. Partial observables are quantities that physical measuring devices can be responsive to, but whose value cannot be predicted

[1] Earman (2002), Earman (2011), and Ruetsche (2014) are notable exceptions. Physicists have been more willing to dive in. Paetz (2010) gives an excellent critical discussion of the many technical challenges faced by the TTH.

given the state alone (e.g., proper time along a worldline). A full observable is understood as a coincidence or correlation of partial observables whose value can be predicted given the state (e.g., proper time along a worldline at the point where it intersects another worldline). Only measurements of full observables can be directly compared to the predictions made by the mechanical theory.

A timeless mechanical system is given by a triple $(\mathcal{C}, \Gamma, f)$. $\mathcal{C}$ is the configuration space of partial observables, q^a. A *motion* of the system is given by an unparametrized curve in $\mathcal{C}$, representing a sequence of correlations between partial observables. The space of motions, Γ is the statespace of the system and is typically presymplectic. The evolution equation is given by $f = 0$, where f is a map $f : \Gamma \times \mathcal{C} \to V$, and V is a vector space. For systems that can be modeled using Hamiltonian mechanics, Γ and f are completely determined by a surface Σ in the cotangent bundle $T^*\mathcal{C}$ (the space of partial observables and their conjugate momenta p_a). This surface is defined by the vanishing of some Hamiltonian function $H : T^*\mathcal{C} \to \mathbb{R}$.

If the system has a preferred external time variable, the Hamiltonian can be decomposed as

$$H = p_t + H_0(q^i, p_i, t) \tag{1}$$

where t is the partial observables in $\mathcal{C}$ that corresponds to time. Generally covariant mechanical systems lack such a canonical decomposition. Although these systems are fundamentally timeless, it is possible for a notion of time to emerge thermodynamically. A closed system left to thermalize will eventually settle into a time-independent equilibrium state. Viewed as part of a definition of equilibrium, this thermalization principle requires an antecedent notion of time. The TTH inverts this definition and use the notion of an equilibrium state to select a partial observable in $\mathcal{C}$ as time.

Three hurdles present themselves. The first is providing a coherent mathematical characterization of equilibrium states. The second is finding a method for extracting information about the associated time flow from a specification of the state. Finally, in order to count as an emergent explanation of time, one has to show that the partial observable selected behaves as a traditional time variable in relevant limits.

For generally covariant quantum theories, Connes and Rovelli (1994) propose a concrete strategy to overcome these hurdles. Minimally, such a theory can be thought as a non-commutative C^*-algebra of diffeomorphism-invariant

observables, $\mathfrak{A}$, along with a set of physically possible states, $\{\phi\}$.[2] Via the Gelfand-Nemark-Segal (GNS) construction, each state determines a concrete Hilbert space representation $(\pi_\phi(\mathfrak{A}), \mathcal{H}_\phi)$, and a corresponding von Neumann algebra $\pi_\phi(\mathfrak{A})''$, defined as the double commutant of $\pi_\phi(\mathfrak{A})$.

Connes and Rovelli first appeal to the well-known *Kubo-Martin-Schwinger (KMS) condition* to characterize equilibrium states. A state, ρ, on a von Neumann algebra, $\mathfrak{M}$, satisfies the KMS condition for inverse temperature $0 < \beta < \infty$ with respect to a 1-parameter group of automorphisms, $\{\alpha_t\}$, if for any $A, B \in \mathfrak{M}$ there exists a complex function $F_{A,B}(z)$, analytic on the strip $\{z \in \mathbb{C} | 0 < \mathrm{Im} z < \beta\}$ and continuous on the boundary of the strip, such that

$$\begin{aligned} F_{A,B}(t) &= \rho(\alpha_t(A)B) \\ F_{A,B}(t + i\beta) &= \rho(B\alpha_t(A)) \end{aligned} \tag{2}$$

for all $t \in \mathbb{R}$. The KMS condition generalizes the idea of an equliibirum state to quantum systems with infinitely many degrees of freedom. KMS states are stable, passive, and invariant under the dynamics, $\{\alpha_t\}$. Moreover in the finite limit, the KMS condition reduces to the standard Gibbs postulate.

Although the KMS condition is framed relative to a chosen background dynamics, according to the main theorem of *Tomita-Takesaki modular theory*, every faithful state determines a canonical 1-parameter group of automorphisms according to which it is a KMS state. Connes and Rovelli go on to identify the flow of time with the flow of this state-dependent *modular automorphism group*.

In the GNS representation $(\pi_\phi(\mathfrak{A}), \mathcal{H}_\phi)$, the defining state, ϕ, is represented by a cyclic vector $\Phi \in \mathcal{H}_\phi$. If ϕ is a *faithful* state (i.e., if $\phi(A^*A) = 0$ entails that $A = 0$) then the vector Φ is also separating. In this setting we can apply the tools of Tomita-Takesaki modular theory. The main theorem asserts the existence of two unique modular invariants, an antiunitary operator, J, and a positive operator, Δ. (Here we will only be concerned with the latter.) The 1-parameter family, $\{\Delta^{is} | s \in \mathbb{R}\}$, forms a strongly continuous unitary group,

$$\sigma_s(A) := \Delta^{is} A \Delta^{-is} \tag{3}$$

for all $A \in \pi(\mathfrak{A})''$, $s \in \mathbb{R}$. The defining state is invariant under the flow of the modular automorphism group, $\phi(\sigma_s(A)) = \phi(A)$. Furthermore, $\phi(\sigma_s(A)B) =$

[2] See Brunetti et al. (2003) for a formal development of this basic idea.

$\phi(B\sigma_{s-i}(A))$. Thus ϕ satisfies the KMS condition relative to $\{\sigma_s\}$ for inverse temperature $\beta = 1$.

For any faithful state, this procedure identifies a partial observable, the thermal time, $t_\phi := s$, parametrizing the flow of the (unbounded) thermal hamiltonian $H_\phi := -\ln \Delta$, which has Φ as an eigenvector with eigenvalue zero. We can then go on to decompose the timeless Hamiltonian $H = p_{t_\phi} + H_\phi$. Associated with any such state, there is a natural "flow of time" according to which the system is in equilibrium. But in what sense does this thermal time flow correspond to various notions of physical time? In particular, how is thermal time related to the proper time measured by a localized observer?

Although they do not establish a general theorem linking thermal time to proper time, Connes and Rovelli do make substantial progress on the third hurdle in one intriguing special case. For a uniformly accelerating, immortal observer in Minkowski spacetime, the region causally connected to her worldline is the *Rindler wedge*. In standard coordinates we can explicitly write the observer's trajectory as

$$\begin{aligned} x^0(\tau) &= a^{-1}\sinh(\tau) \\ x^1(\tau) &= a^{-1}\cosh(\tau) \\ x^2(\tau) &= x^3(\tau) = 0 \end{aligned} \tag{4}$$

where τ is the observer's proper time. The wedge region is defined by the condition $x^1 > |x^0|$. The *Bisognano-Wichmann theorem* then tells us that in the vacuum state, the modular automorphism group for the wedge implements wedge-preserving Lorentz boosts — Δ^{is} is given by the boost $U(s) = e^{2\pi i s K_1}$ (where K_1 is the representation of the generator of an x^1-boost). Since the Lorentz boost $\lambda(a\tau)$ implements a proper time translation along the orbit of an observer with acceleration a, $U(\tau) = e^{ai\tau K_1}$ can be viewed as generating evolution in proper time. Comparing these two operators, we find that proper time is directly proportional to thermal time,

$$s = \frac{2\pi}{a}\tau \tag{5}$$

The Unruh temperature measured by the observer is $T = a/2\pi k_b$ (where k_b is Boltzmann's constant), this leads Connes and Rovelli to propose that the Unruh temperature can be interpreted as the ratio between thermal and proper time. Not only does this relationship hold along the orbits of constant

acceleration, but if an observer constructs global time coordinates for the wedge via the process of Einstein synchronization, this global time continues to coincide with the rescaled thermal time flow.

We can now summarize the main content of the TTH:

Thermal Time Hypothesis (Rovelli-Connes)**.** *In a generally covariant quantum theory, the flow of time is defined by the state-dependent modular automorphism group. The Unruh temperature measured by an accelerating observer represents the ratio between this time and her proper time.*

This is a bold idea with a numerous potential implications for quantum physics and cosmology. Over the next three sections, we will consider a series of technical and conceptual objections to the TTH.

3 Thermal Time and Proper Time

The Bisognano Wichmann theorem only applies to immortal, uniformly accelerating observers in the vacuum state of a quantum field theory in flat spacetime. How can we characterize the relationship between thermal and proper time for a broader, more physically realistic class of observers and theories?

A uniformly accelerating mortal observer has causal access to a different region of Minkowski spacetime, the *doublecone* formed by the intersection of her future lightcone at birth and her past lightcone at death. Because wedges and doublecones can be related by a conformal transformation, in conformally invariant theories, geometric results from wedge algebras can be transferred onto the doubelcone algebras. In the vacuum state of a conformal theory, the doublecone modular automorphism group acts as Hislop-Longo transformations (Hislop and Longo, 1982). Martinetti and Rovelli (2003) use this result to calculate the corresponding relationship between thermal time and proper time for a uniformly accelerating mortal observer:

$$s = \frac{2\pi}{La^2}(\sqrt{1 + a^2L^2} - \cosh a\tau) \tag{6}$$

where L is the observer's lifetime. (The relationship is more complicated in this case due to the fact that proper time is bounded while modular time is unbounded.) For most of the observer's lifespan, s is an approximately constant function of τ, allowing the Unruh temperature to again be interpreted as the local ratio between thermal and proper time.

This is the best we can hope for. Trebels (1997) proves that arbitrary doublecone automorphisms act as local dynamics, only if they act as scaled Hislop-Longo transformations.[3] Of course, if nature is described by a non-conformal theory, then there is no guarantee that the doublecone modular automorphisms will have a suitable geometric interpretation. Saffary (2005) goes further, arguing that they will not have geometric significance in any theory with massive particles. The mathematical results backing this conjecture, however, are only partial.[4]

Attempting to generalize the TTH to cover non-uniform acceleration and non-vacuum states generates further difficulties. Work on the Unruh effect for non-uniformly accelerating observers (e.g., Jian-yang et al. 1995), indicates that such observers feel an acceleration-dependent thermal bath, reflecting the shifting ratio between constant thermal time and acceleration-dependent proper time. The TTH must explain the phenomenological experience of the observer who will presumably age according to her proper time, not the background thermal time flow. On top of this, if the global state is not a vacuum state, then it is not clear that the wedge modular automorphisms will carry a dynamical interpretation at all. The Radon-Nikodym theorem ensures that the action of the modular automorphism group uniquely determines the generating state. If ϕ, ψ are two (faithful, normal) states on a von Neumann algebra $\mathfrak{M}$, then the associated modular automorphism groups σ^t_ϕ, σ^t_ψ differ by a non-trivial inner automorphism, $\sigma^t_\phi(A) = U\sigma^t_\psi(A)U^*$, for all $A \in \mathfrak{M}$, $t \in \mathbb{R}$, so the general wedge dynamics will not be simple rescalings of the vacuum case.

None of these are knockdown objections since so little is known about the geometric action of modular operators apart from the Bisognano-Wichmann theorem and its conformal generalization. But our current ignorance also presents a major challenge. (The situation is even less clear in general curved spacetime settings.) The defender of the TTH has at least four options on

[3]Formally, Trebels requires that local dynamics be continuous 1-parameter groups of automorphisms of the doublecone algebra that preserve subalgebra localization as well as spacelike and timelike relations between interior points. For a detailed discussion of Trebels's results, see Borchers (2000), §3.4.

[4]In the massless case, the modular generators are ordinary differential operators, δ_0, of order 1. In the massive case, it has been conjectured that the modular generators are pseudo-differential operators $\delta_m = \delta_0 + \delta_r$, where the leading term is given by the massless generator δ_0 and δ_r is a pseudo-differential operators of order < 1. This second term is thought to give rise to non-local action without geometric interpretation.

the table.

She can hold out hope for a suitably general dynamical interpretation of modular automorphisms in a wide class of physically significant states. There is some indication that states of compact energy (e.g., states satisfying the Döplicher-Haag-Roberts and Buchholz-Fredenhagen selection criteria) give rise to well-behaved modular structure on wedges. In this case the wedge modular automorphisms can be related to those in the vacuum state by the Radon-Nikodym derivative (Borchers, 2000). The analogous problem for doublecones is still open.

Alternatively, she could reject the idea that the thermal time flow determines the temporal metric directly. Thermal time would only give rise to the order, topological, and group theoretic properties of physical time. Metrical properties would be determined by a completely different set of physical relations. Some support for this idea comes from the justification of the clock hypothesis in general relativity. Rather than stipulating the relationship between proper time, τ, and the length of a timelike curve $||\gamma||$, Fletcher (2013) shows that for any $\epsilon > 0$, there is an idealized lightclock moving along the curve which will measure $||\gamma||$ within ϵ. This justifies the clock hypothesis by linking the metrical properties of spacetime to the readings of tiny lightclocks. If the metrical properties of time experienced by localized observers arises via some physical mechanism akin to light clock synchronization. This would explain why the duration of time felt by the observer matches her proper time and not the geometrical flow of thermal time.

Perhaps motivated by the justification of the clock hypothesis, the defender of the TTH could attempt to argue that the metrical properties of time emerge from modular dynamics in the short distance limit of the theory. If the theory has a well-defined ultraviolet limit, the renormalization group flow should approach a conformal fixed point. Buchholz and Verch (1995) prove that in this limit, the double-cone modular operators act geometrically like wedge operators implementing proper time translations along the observer's worldline. It is unlikely that the physics at this scale would directly impact phenomenology, but the asymptotic connection might turn out to be important for explaining the metrical properties of spacetime (which bigger, more realistic lightclocks measure) as emergent features of some underlying theory of quantum gravity.

A final option would be to go back to the drawing board. Rovelli and Connes briefly note that since the modular automorphisms associated with each (faithful, normal) state of a von Neumann algebra are connected by

inner automorphisms, they all project down onto the same 1-parameter group of outer automorphisms the algebra. The TTH could be revised to claim that this canonical state-independent flow represents the non-metrical flow of physical time. It is not known, however, under what circumstances the outer flow acts in suitably geometric fashion to be interpretable as local dynamics, so it remains to be seen whether or not this is a viable option. The move does have immediate consequences for the global dynamics, however. Since the global algebra is expected to be type I, all modular automorphisms will be inner. As a result the canonical group of outer automorphisms is trivial. At a global level, there is no passage of time. At the local level, time emerges as a consequence of our ignorance of the global state.

4 The Classical Limit

The classical limit presents a different kind of challenge. Conceptually, nothing about the idea that a statistical state selects a preferred thermal time requires that the theory be quantum mechanical. The proposed mechanism for selecting a partial observable using modular theory, however, does appear to rely on the noncommutativity of quantum observables. If we model classical systems using abelian von Neumann algebras, then every state is tracial (i.e., $\phi(AB) = \phi(BA)$), and consequently every associated modular automorphism group acts as the identity, trivializing the thermal time flow. Does the TTH have a classical counterpart, or is quantum mechanics required to save time in a generally covariant setting?

Arguing by analogy with standard quantization procedures, Connes and Rovelli suggest that in the classical limit commutators need to be replaced by Poisson brackets. We begin with an arbitrary statistical state, ρ, represented by a probability distribution over a classical statespace Γ:

$$\int_{\Gamma} dx\ \rho(x) = 1 \tag{7}$$

where $x \in \Gamma$ is a timeless microstate. By analogy with the Gibbs postulate, we can introduce the "thermal Hamiltonian,"

$$H_{\rho} = -\ln \rho \tag{8}$$

With respect to the corresponding Hamiltonian vector field, the evolution of

an arbitrary classical observable, $f \in C^{\infty}(\Gamma)$, is given by

$$\frac{d}{ds} f = \{-\ln \rho, f\} \tag{9}$$

and $\rho = \exp(-H_{\rho})$. With respect to the Poisson bracket structure, the classical algebra of observables is non-abelian. Gallavotti and Pulvirenti (1976) use this non-abelian structure to define an analogue of the KMS condition. Is this connection strong enough to support a version of the TTH in ordinary general relativity? Or does it only serve to aid us in understanding how the thermal time variable behaves in the transition from quantum theory to classical physics?

The difficulty lies in connecting the thermal time flow for an arbitrary statistical state to our ordinary conception of time. In the quantum case this link was provided by the Bisognano-Wichmann theorem, which does not have a classical analogue. The problem is magnified by the lack of a full understanding of statistical mechanics and thermodynamics in curved spacetime. Rovelli has done some preliminary work on developing a full theory of generally covariant thermodynamics based on the foundation supplied by the TTH, including an elegant derivation of the Tolman-Ehrenfest effect, but the field is still young.[5]

Setting aside these broader interpretive challenges for now, an important first step lies in obtaining a better understanding the classical selection procedure outlined above. As it turns out, the commutator-to-Poisson-bracket ansatz is on firmer foundational footing than one might initially suspect. As emphasized by Alfsen and Shultz (1998), non-abelian C^*-algebras have a natural *Lie-Jordan structure*:

$$AB = A \bullet B - i(A \star B) \ , \tag{10}$$

The non-associative Jordan product, $\bullet$, encodes information about the spectra of observables, while the associative Lie product, $\star$, encodes the generating relation between observables and symmetries. The significance of the commutator, is that it defines the canonical Lie product, $A \star B := i/2[A, B]$. Classical mechanical theories formulated on either a symplectic or Poisson manifold have a natural Lie-Jordan structure as well. The standard product of functions defines an associative Jordan product, encoding spectral information, while the Poisson bracket determines the associative Lie product,

[5]See Rovelli and Smerlak (2011).

describing how classical observables generate Hamiltonian vector fields on statespace. Together, this structure is called a *Poisson algebra*. The primary difference between the classical and quantum cases is the associativity/non-associativity of the Jordan product.

These considerations point towards the idea that the appropriate classical analogue of a noncommutative von Neumann algebra, is not a commutative von Neumann algebra, but a Poisson algebra. In this setting, initial strides towards a classical analogue of modular theory have been made by Weinstein (1997). Given any smooth density, μ, on a Possion manifold, Γ, Weinstein defines a corresponding *modular vector field* ϕ_μ given by the operator $\phi_\mu : f \to \mathrm{div}_\mu H_f$ where H_f is the Hamiltonian vector field associated with a classical observable, $f \in C^\infty(\Gamma)$. The antisymmetry of the Poisson bracket entails that the operator ϕ_μ is a vector field on Γ. Weinstein proposes ϕ_μ as the classical analogue of the modular automorphism group. It characterizes the extent to which the Hamiltonian vector fields are divergence free (with respect to the density μ), vanishing iff all Hamiltonian vector fields are divergence free.

We can connect Weinstein's classical modular theory to the TTH. If Γ is a symplectic manifold and we let μ be the density associated with the canonical Liouville volume form, then $\phi_\mu(f) = 0$ for all observables. This reflects the conservation of energy by Hamiltonian flows in symplectic dynamical systems. Given any statistical state, however, we can define an associated density which leads to a nontrivial modular vector field. For any positive function, h, we have

$$\phi_{h\mu} = \phi_\mu + H_{-\ln h} = H_{-\ln h}. \tag{11}$$

Therefore any statistical state, ρ, defines a modular vector field equivalent to the Hamiltonian vector field $H_{-\ln \rho}$ associated with the density $e^{-\ln \rho}\mu$. We immediately recognize $-\ln \rho$ as the thermal Hamiltonian postulated by Connes and Rovelli. Clearly, $e^{is \ln \rho} \rho e^{-is \ln \rho} = \rho$, thus the state is invariant with respect to the flow of $H_{-\ln \rho}$. Additionally, it can be shown that ρ satisfies the KMS condition with respect to these dynamics, hence, from the perspective of the associated time flow ρ resembles an invariant equilibrium state just as in the quantum case.

5 Conceptual Challenges

As we have seen in the previous two sections, the TTH faces a number of technical challenges (some of which look easier to overcome than others). There are, however, several deeper conceptual problems looming in the background which pose a more serious challenge to the viability of the hypothesis. Here, we will discuss two of the most pressing.

The first, which we will call the *generality problem*, draws upon the preceding discussion of the classical limit. While mathematically speaking, Weinstein's modular vector field gives us a method for selecting a canonical thermal time flow in a classical theory, physical speaking, there is no reason why we should view the corresponding thermal time as physical time. As we have seen, any statistical state determines thermal dynamics according to which it is a KMS state, however, if ρ is a non-equilibrium state, the resultant thermal time flow does not align with our ordinary conception of time. By the lights of thermal time, a cube of ice in a cup of hot coffee is an invariant equilibrium state! The same problem arises in the quantum domain — only for states which are true equilibrium states will the thermal time correspond to physical time.

It appears inevitable that the TTH will have to be tempered. Rather than letting any state determine a corresponding flow of thermal time, only certain reference states should be permitted. Apart from the problem of providing an intrinsic, non-dynamical characterization of such states, if a system is not in one of these, it is hard to envision how a counterfactual state of affairs can determine the actual flow of time.[6] This might provide more reasons for the defender of the TTH to explore the state independent, outer modular flow. Alternatively, she could try to argue that local non-equilibrium behavior can be viewed as small fluctuations from some background state. On this approach, the local flow of time in my office according to which the ice

[6] A closely related worry, what we might call the *background-dependence problem*, has been voiced by Earman (2011) and Ruetsche (2014). Their concern is that we can only identify modular automorphisms as dynamics because we already have a rich spatiotemporal geometry in the background. This casts doubt on whether the TTH can provide a coherent definition of time in situations where such structure is absent (as required to solve the full problem of time). This is exacerbated if the TTH is modified in response to the generality problem. Unless the modular automorphism group can always be viewed dynamically, the defender of the TTH will be hard pressed to find constraints capable of separating the dynamical cases from the non-dynamical cases which are independent of all background temporal structure.

melts and the coffee cools is not defined by the thermal state of the ice/coffee system, but the thermal state of some larger enveloping system (the entire universe perhaps). Rovelli (1993) hints in this direction, calculating that in a Friedman-Robertson-Walker universe, the thermal time induced by the equilibrium state of the cosmic microwave background will be proportional to the FRW time. While the connection is intriguing, it seems unlikely that an explanation of this sort will be able to account for the flow of time experienced by localized, mortal observers like us. It would be truly remarkable to discover that our faculties of perception are sensitive to the thermal features of the CMB.

The second problem is the *gauge problem*. The TTH does succeed in providing a means to select a privileged 1-parameter flow on the space of full, gauge invariant observables of a generally covariant theory. What makes this flow interpretable as a *dynamical* flow, however, is its description as a sequence of correlations between partial observables. The difficulty is that these partial observables are not diffeomorphism invariant. Assuming that we treat diffeomorphisms in generally covariant theories as standard gauge symmetries (which is how we got into the problem of time in the first place), then the partial observables are just descriptive fluff. They do not directly represent physical features of our world.

The problem is *not* the resultant timelessness of fundamental physics. The TTH adopts this dramatic conclusion willingly. The problem is that the TTH is supposed to explain how the appearance of time and change emerge from timeless foundations. But the explanation given is couched in gauge-dependent language, and it is not apparent how we can extract a gauge invariant story from it. We can introduce partial observables and use correlations between them to calculate and predict emergent dynamical behavior, but we cannot use these correlations to *explain* that behavior. We lack a gauge invariant picture of generally covariant theories, and the TTH, at least in its present form, does not provide one.

Can a revised TTH give us the explanatory tools needed to understand the flow of time without reference to partial observables, or, does the entire framework of timeless mechanics require us to revise our conception of how ontology, explanation, and gauge symmetries are related?[7] Whether or not

[7]Drifting in the latter direction, Rovelli (2014) suggests that gauge-dependent quantities are more than just mathematical redundancies, "they describe handles through which systems couple: they represent real relational structures to which the experimentalist has access in measurement by supplying one of the relata in the measurement procedure itself."

quantum thermodynamics can save time may rest on the solutions to these new incarnations of vexingly familiar philosophical problems.

References

Alfsen, E. and F. Shultz (1998). Orientation in operator algebras. *Proceedings of the National Academy of Sciences, USA 95*, 6596–6601.

Borchers, H. J. (2000). On revolutionizing quantum field theory with Tomita's modular theory. *Journal of Mathematical Physics 41*(6), 3604–3673.

Brunetti, R., K. Fredenhagen, and R. Verch (2003). The generally covariant locality principle – a new paradigm for local quantum field theory. *Communications in Mathematical Physics 237*, 31–68.

Buchholz, D. and R. Verch (1995). Scaling algebras and renormalization group in algebraic quantum field theory. *Reviews in Mathematical Physics 7*, 1195.

Connes, A. and C. Rovelli (1994). Von Neumann algebra automorphisms and time-thermodynamics relation in generally covariant quantum theories. *Classical and Quantum Gravity 11*(12), 2899.

Earman, J. (2002). Thoroughly modern McTaggart. *Philosopher's Imprint*, 2. http://www.philosophersimprint.org/002003/.

Earman, J. (2011). The Unruh effect for philosophers. *Studies in History and Philosophy of Modern Physics 42*, 81–97.

Fletcher, S. (2013). Light clocks and the clock hypothesis. *Foundations of Physics 43*, 1369–1383.

Gallavotti, G. and M. Pulvirenti (1976). Classical KMS condition and Tomita-Takesaki theory. *Communications in Mathematical Physics 46*, 1–9.

Hislop, P. D. and R. Longo (1982). Modular structure of the local algebras associated with a free massless scalar field theory. *Communications in Mathematical Physics 84*, 71.

Jian-yang, Z., B. Aidong, and Z. Zheng (1995). Rindler effect for a nonuniformly accelerating observer. *International Journal of Theoretical Physics 34*, 2049–2059.

Martinetti, P. and C. Rovelli (2003). Diamond's temperature: Unruh effect for bounded trajectories and thermal time hypothesis. *Classical and Quantum Gravity 20*(22), 4919.

Paetz, T.-T. (2010). An analysis of the 'thermal-time concept' of Connes and Rovelli. Master's thesis, Georg-August-Universität Göttingen.

Rovelli, C. (1993). The statistical state of the universe. *Class.Quant.Grav. 10*, 1567.

Rovelli, C. (2011). Forget time: Essay written for the FQXi contest on the nature of time. *Foundations of Physics*.

Rovelli, C. (2014). Why gauge? *Foundations of Physics 44*(1), 91–104.

Rovelli, C. and M. Smerlak (2011). Thermal time and Tolman–Ehrenfest effect: 'temperature as the speed of time'. *Classical and Quantum Gravity 28*(7), 075007.

Ruetsche, L. (2014). Warming up to thermal the thermal time hypothesis. Quantum Time Conference, University of Pittsburgh, March 28-29.

Saffary, T. (2005). *Modular Action on the Massive Algebra*. Ph. D. thesis, Hamburg.

Trebels, S. (1997). *Über die Geometrische Wirkung Modularer Automoprhismen*. Ph. D. thesis, Göttingen.

Weinstein, A. (1997). The modular automorphism group of a Poisson manifold. *Journal of Geometry and Physics 23*, 379–394.

Neural redundancy and its relation to neural reuse

Abstract

Evidence of the pervasiveness of neural reuse in the human brain has forced a revision of the standard conception of modularity in the cognitive sciences. One persistent line of argument against such revision, however, draws from a large body of experimental literature attesting to the existence of cognitive dissociations. While numerous rejoinders to this argument have been offered over the years, few have grappled seriously with the phenomenon. This paper offers a fresh perspective. It takes the dissociations seriously, on the one hand, while affirming that traditional modularities of mind do not do justice to the evidence of neural reuse, on the other. The key to the puzzle is neural redundancy. The paper offers both a philosophical analysis of the relation between reuse and redundancy, as well as a plausible solution to the problem of dissociations.

1. Introduction

Cognitive science, linguistics and the philosophy of psychology have long been under the spell of "the modularity of mind" (Fodor 1983), or the idea of the mind as a modular system (see e.g. de Almeida and Gleitman 2018). In contemporary psychology, a modular system is generally understood to be "one consisting of functionally specialized subsystems responsible for processing different classes of input (e.g. for vision, hearing, human faces, etc.), or at any rate for handling specific cognitive tasks" (Zerilli 2017a, 231). According to this theory, "human cognition can be decomposed into a number of functionally independent processes, [where] each of these processes operates over a distinct domain of cognitive information" (Bergeron 2007, 176). What makes one process distinguishable from another is its "functional independence, the fact that one can be affected, in part or in totality, without the other being affected, and vice versa" (Bergeron 2007, 176). Furthermore, given that functional processes are realized in the brain, a functionally specialized process is one which presumably occupies a distinctive portion of neural tissue, though not necessarily a small, closely circumscribed and contiguous region. So fruitful and influential has this model been that it is safe to say that in many quarters of the cognitive sciences—and most especially in cognitive psychology, cognitive neuropsychology and evolutionary psychology—modularity is essentially the received view (McGeer 2007; Carruthers 2006; de Almeida and Gleitman 2018).

Developments in cognitive neuroscience over the past thirty years, however, have discomfited the modular account. More evidence than ever before points to the pervasiveness of neural reuse in the human brain—the "redeployment" or "recycling" of neural circuits over widely disparate cognitive domains (Anderson, 2010, 2014; Dehaene, 2005). As the terminology suggests, theories of "re-use" posit the "exaptation" of established and diachronically stable neural circuits over the course of evolution or normal development *without* loss of original function, so that the functional contribution of a circuit is preserved across multiple task domains.[1] As Anderson (2010, 246) explains, "rather than posit a functional architecture for the brain whereby individual regions are dedicated to large-scale cognitive domains like vision, audition, language and the like, neural reuse theories suggest that low-level neural circuits are used and reused for various purposes in different cognitive and task domains." According to the theory, just the same circuits exapted for one purpose can be exapted for another provided sufficient intercircuit pathways exist to allow alternative arrangements of them. Indeed, the same parts put together in *different* ways will yield different functional outcomes, just as "if one puts together the same parts *in the same way* one will get the same functional outcomes" (Anderson 2010, 247, my emphasis). The evidence here converges from heterogeneous sources and research paradigms, including neuroimaging (Anderson 2007a; 2007b; 2007c; 2008), computational (Eliasmith 2015), biobehavioral (Casasanto and Dijkstra 2010) and interference paradigms (Gauthier et al.

[1] This usage of "exaptation" is somewhat misleading, since exaptation usually implies loss of original function (see Godfrey-Smith 2001).

2003), and exempts practically no area of the brain (Leo et al. 2012, 2), including areas long regarded as specialized hubs for certain types of sensory processing, e.g. visual and auditory pathways (Striem-Amit and Amedi 2014). Among other things, this means that one of the hallmark features of a module—its domain specificity (Coltheart 1999)—looks too stringent a requirement to prove useful.[2] For neural reuse demonstrates that any one module will typically be sensitive to *more* than one stimulus, including—most importantly—those channeled along intermodal pathways. Meanwhile efforts to salvage a computational or "software" theory of modularity, which carries no commitments regarding implementation, have met with scepticism (Anderson 2007c; 2010; Anderson & Finlay 2014) if not outright opposition (Zerilli 2017a).[3] And while the brain could still be modular in some other sense, what is clear is that the strict domain-specific variety of modularity can no longer serve as an appropriate benchmark.[4]

And yet there is a persistent line of argument *against* this conclusion which draws from a large body of experimental literature attesting to the existence of cognitive

[2] The sense of domain specificity that is relevant here refers to a module's sensitivity to a restricted class of inputs as defined by a domain of psychology—such as visual, auditory or linguistic information. For discussion of alternative senses, see Barrett and Kurzban (2006) and Prinz (2006).

[3] Though by no means universally (see e.g. Carruthers 2010; Jungé and Dennett 2010).

[4] Nor, for that matter, can its cognate property, informational encapsulation (see below).

dissociations, in which a cognitive ability (say language) is either selectively impaired (linguistic ability is compromised, but no other cognitive ability seems to be materially affected) or selectively spared (general intelligence is compromised, while linguistic abilities function more or less as they should). This literature, most vividly exemplified in lesion studies, is frequently cited in support of classical modularities of mind—be they inspired by the likes of Jerry Fodor (1983), evolutionary psychology (e.g. Cosmides and Tooby 1994; Barrett and Kurzban 2006; Carruthers 2006) or some variation thereof (e.g. ACT-R). While numerous rejoinders to this line of thinking have been offered over the years, few have grappled seriously with the phenomenon, either dismissing the dissociations as noisy, or reasoning from architectural considerations that even nonmodular systems can generate dissociations (Plaut 1995). The aim of this paper is to offer a fresh perspective on this vexed topic. I take the dissociation evidence seriously, on the one hand, while affirming that traditional modularities of mind do not do justice to the evidence of neural reuse, on the other. I do this by invoking neural redundancy, an important feature of cortical design that ensures we have various copies of the same elementary processing units that can be put to alternative (if computationally related) uses in enabling diverse cognitive functions. In the course of the discussion I offer a philosophical explication of the relationship between neural reuse and neural redundancy.

2. What is the Problem? Cognitive Dissociations and Neural Reuse

Let us take an especially contentious question to underscore the nature of the problem we are dealing with and how redundancy might assist in its illumination. The question is this: Does language rely on specialized cognitive and neural machinery, or does it rely on the same machinery that allows us to get by in other domains of human endeavour? The question is bound up with many other questions of no less importance, questions concerning the uniqueness of the human mind, the course of biological evolution and the power of human culture. What is perhaps a little unusual about this question, however—unusual for a question whose answer concerns both those working in the sciences and the humanities—is that it can be phrased as a polar interrogative, i.e. as a question which admits of a yes or no response. And indeed the question has divided psychologists, linguists and the cognitive science community generally for many decades now, more or less into two camps. I would like to sketch the beginnings of an answer to this question—and others like it—in a way that does not pretend it can receive a simple yes or no response.

First of all, let me stress again that neural reuse is as well verified a phenomenon as one can expect in the cognitive sciences, and that it has left virtually no domain of psychology untouched. Neural reuse suggests that there is nothing so specialized in the cortex that it cannot be repurposed to meet new challenges while retaining its capacity for meeting old ones. In that regard, to be sure, what I am proposing is unapologetically on the side of those who maintain that language, as well as many other psychological capacities, are

not cognitively special—e.g. that there is no domain-specific "language organ" (cf. Chomsky 1980,39, 44; 1988, 159; 2002, 84-86).

And yet I would like to carefully distinguish this claim from the claim that there are no areas of the brain that subserve exclusively linguistic functions. The neuropsychological literature offers striking examples of what appear to be fairly clean dissociations between linguistic and nonlinguistic capacities, i.e. cases in which language processing capacities appear to be disrupted without impeding other cognitive abilities, and cases in which the reverse situation holds (Fedorenko et al. 2011; Hickok and Poeppel 2000; Poeppel 2001; Varley et al. 2005; Luria et al. 1965; Peretz and Coltheart 2003; Apperly et al. 2006). An example would be where the ability to hear words is disrupted, but the ability to recognize non-word sounds is spared (Hickok and Poeppel 2000; Poeppel 2001). Discussing such cases, Pinker and Jackendoff (2005, 207) add that "[c]ases of amusia and auditory agnosia, in which patients can understand speech yet fail to appreciate music or recognize environmental sounds...show that speech and non-speech perception in fact doubly dissociate." Although dissociations are to some extent compatible with reuse—indeed there is work suggesting that focal lesions can produce specific cognitive impairments within a range of nonclassical architectures (Plaut 1995)—and it is equally true that often the dissociations reported are noisy (Cowie 2008), still their very ubiquity needs to be taken seriously and accounted for in a more systematic fashion than many defenders of reuse have been willing to do (see e.g. Anderson 2010, 248; 2014, 46-48). After all, a good deal of support for

theories of reuse comes from the neuroimaging literature, which is somewhat ambiguous taken by itself. As Fedorenko et al. (2011, 16428) explain:

> standard functional MRI group analysis methods can be deceptive: two different mental functions that activate neighbouring but non-overlapping cortical regions in every subject individually can produce overlapping activations in a group analysis, because the precise locations of these regions vary across subjects, smearing the group activations. Definitively addressing the question of neural overlap between linguistic and nonlinguistic functions requires examining overlap within individual subjects, a data analysis strategy that has almost never been applied in neuroimaging investigations of high-level linguistic processing.

When Fedorenko and her colleagues applied this strategy themselves, they found that "most of the key cortical regions engaged in high-level linguistic processing are not engaged by mental arithmetic, general working memory, cognitive control or musical processing," and they think that this indicates "a high degree of functional specificity in the brain regions that support language" (2011, 16431). While I do not believe that claims of this strength have the least warrant—as I shall explain, functional specificity cannot be established merely by demonstrating that a region is selectively engaged by a task—these results do at least substantiate the dissociation literature in an interesting way and make it more difficult for

those who would prefer to dismiss the dissociations with a ready-made list of alternative explanations. Similar results were found by Fedorenko et al. (2012).

3. How Might Redundancy Feature In a Solution?

With rare exceptions (e.g. Friston and Price 2003; Barrett and Kurzban 2006; Jungé and Dennett 2010), redundancy has passed almost unnoticed in the philosophical and cognitive science literature. This is in stark contrast to the epigenetics literature, where redundancy and the related concept of degeneracy[5] have been explored to some depth (e.g. see Edelman and Gally 2001; Mason 2010; Whiteacre 2010; Deacon 2010; Iriki and Taoka 2012; Maleszka et al. 2013). The idea behind neural redundancy is that, for good evolutionary reasons (see below), the brain incorporates a large measure of redundancy of function. Brain regions (such as cortical columns and similar structures) fall in an iterative, repetitive and almost lattice-like arrangement in the cortex. Neighbouring columns have similar response properties: laminar and columnar changes are for the most part smooth—not abrupt—as one moves across the cortex, and adjacent modules do not differ markedly from one another in their basic structure and computations (if they really differ at all when taken in such

[5] Redundancy occurs when items have the same structure and function (i.e. are both isomorphic and isofunctional). Degeneracy occurs when items having *different* structures can perform the same function (i.e. are heteromorphic but isofunctional). Degeneracy implies genuine multiple realization (see Zerilli 2017b).

proximity). Regional *solitarity* is therefore not likely to be a characteristic of the brain (Anderson 2014, 141).[6] That is to say, we do not possess just one module for X, and one module for Y, but in effect several *copies* of the module for X, and several copies of the module for Y, all densely stuffed into the same cortical zones. As Buxhoeveden and Casanova (2002, 943) explain of neurons generally:

> In the cortex, more cells do the job that fewer do in other regions....As brain evolution paralleled the increase in cell number, a reduction occurred in the sovereignty of individual neurones; fewer of them occupy critical positions. As a consequence, plasticity and redundancy have increased. In nervous systems containing only a few hundred thousand neurones, each cell plays a more essential role in the function of the organism than systems containing billions of neurones.

The same principle very likely holds for functionally distinct groupings of neurons (i.e. cortical columns and like structures), as Jungé and Dennett (2010, 278) conjecture:

> It is possible that specialized brain areas contain a large amount of structural/computational redundancy (i.e., many neurons or collections of neurons

[6] The term "solitarity" is Anderson's, but while he concedes that solitarity will be "relatively rare," he does not appear to believe that anything particularly significant follows from this. See also Anderson (2010, 296).

> that can potentially perform the same class of functions). Rather than a single neuron or small neural tract playing roles in many high-level processes, it is possible that distinct subsets of neurons within a specialized area have similar competencies, and hence are redundant, but as a result are available to be assigned individually to specific uses....In a coarse enough grain, this neural model would look exactly like multi-use (or reuse).

This is plausibly why capacities which are functionally very closely related, but which for whatever reason are forced to recruit different neural circuits, will often be localized in broadly the same regions of the brain. For instance, first and second languages acquired early in ontogeny settle down in nearly the same region of Broca's area; and even when the second language is acquired in adulthood the second language is represented nearby within Broca's area (while artificial languages are not) (Kandel & Hudspeth 2013). The neural coactivation graphs of such composite networks must look very similar. Indeed these results suggest—and a redundancy model would predict—that two very similar tasks which are forced to recruit different neural circuits should exhibit similar patterns of activation. And this is more or less what we find (see below).

One might be tempted to think that redundancy and reuse pull in opposite directions. This is because whereas reuse posits that neural circuits get reused across different tasks and task categories, redundancy accommodates the likelihood of diverse

cognitive functions being activated by structurally and computationally equivalent circuits running in parallel: instead of a single circuit being reused across domains, two, three or more *copies* of that same circuit may be recruited differentially across those domains, such that no *single* circuit gets literally "re-used." But there is no substantive tension here. The redundancy account in truth *supplements* the reuse picture in a way that is consistent with the neuroimaging data, faithful to the core principle of reuse, and compatible with the apparent modularization and separate modifiability of technical and acquired skills in ontogeny. Evidence of the reuse of neural circuits to accomplish different tasks has, in fact, been adduced in aid of a theory which posits the reuse of the same neural *tokens* to accomplish these different tasks. Redundancy means we must accept that at least some of the time what we may actually be witnessing is reuse of the same *types* to accomplish these tasks. This does not diminish the standing of reuse. Let me explain.[7]

To the extent that a particular composite reuses types, and is dissociable pro tanto—residing in segregated brain tissue that is not active outside the domain in question—it is true that to that extent its constituents will *appear* to be domain-specific. But in this case looks will be deceiving. The classical understanding of domain specificity in effect *assumes* solitarity—that a module for X does something which no other module can do as well, or

[7] For a developmental twist on the type/token distinction invoked in the context of modular theorizing about the mind, see Barrett (2006).

that even *if* another module can do X as well, taken together these X-ing modules do not perform outside the X-domain. Here is an example of the latter idea (Bergeron 2007, 176):

> a pocket calculator could have four different division modules, one for dividing numbers smaller than or equal to 99 by numbers smaller than or equal to 99, a second one for dividing numbers smaller than or equal to 99 by numbers greater than 99, a third one for dividing numbers greater than 99 by numbers greater than 99, and a fourth one for dividing numbers greater than 99 by numbers smaller than or equal to 99. In such a calculator, these four capacities could all depend on (four versions of) the same algorithm. Yet, random damage to one or more of these modules in a number of such calculators could lead to observable (double) dissociations between any two of these functions.

Here, each module performs fundamentally the same algorithm, but in distinct hardware, such that dissociations are observable between any two functions. Notice, however, that none of these modules performs outside the "division" domain. This is what allows such duplicate modules to be considered domain-specific—they perform functions which, for all that they might run in parallel on duplicate hardware, are unique to a specific domain of operation, in this case division. If such modules could do work outside the division domain, they would lose the status of domain specificity, and acquire the status of domain neutrality (i.e. they would be domain-general). This is why a module that appears dedicated to a

particular function may not be domain-specific in the classical sense. Dedication is not the same as domain specificity, and redundancy, whether of calculator algorithms or neural circuits, explains why. A composite of neural regions will be dedicated without being domain-specific if its functional resources are accessible to other domains through the deployment (reuse) of neural surrogates (i.e. redundant or "proxy" tokens). In this case its constituents will be multi-potential but single-use (Jungé & Dennett 2010, 278), and the domain specificity on display somewhat cosmetic. To take an example with more immediate relevance to the brain, a set of cortical columns that are structurally and computationally similar may be equally suited for face recognition tasks, abstract-object recognition tasks, the recognition of moving objects, and so on. One of these columns could be reserved for faces, another for abstract objects, another for moving objects, and so on. What is noteworthy is that while the functional activation may be indistinguishable in each case, and the same *type* of resource will be employed on each occasion, a different *token* module will be at work at any one time. To quote Jungé and Dennett (2010, 278) again:

> In an adult brain, a given neuron [or set of neurons] would be aligned with only a single high-level function, whereas each area of neurons would be aligned with very many different functions.

Such modules (and composites) are for all intents and purposes *qualitatively* identical, though clearly not *numerically* identical, meaning that while they share their properties, they

are not *one and the same* (Parfit 1984). The evidence of reuse is virtually all one way when it comes to the pervasiveness of functional inheritance across cognitive domains. It may be that this inheritance owes to reuse of the same tokens (literal reuse) or to reuse of the same types (reuse by proxy), but the inheritance itself has been amply attested. This broader notion of reuse still offers a crucial insight into the operations of cognition, and I dare say represents a large part of the appeal of the original massive redeployment hypothesis (Anderson 2007c).

It is interesting to note in this respect that although detractors have frequently pointed out the ambiguity of neuroimaging evidence on account of its allegedly coarse spatial resolution (e.g. Carruthers 2010), suggesting that the same area will be active across separate tasks and task categories even if distinct but spatially adjacent and/or interdigitated circuits are involved in each case, this complaint can have no bearing on reuse by proxy. Fedorenko et al. (2011, 16431) take their neuroimaging evidence to support "a high degree of functional specificity in the brain regions that support language," but their results do not license this extreme claim. The regions they found to have been selectively engaged by linguistic tasks were all adjacent to the regions engaged in nonlinguistic tasks. Elementary considerations suggest that they have discovered a case of reuse by proxy involving language: the domains tested (mental arithmetic, general working memory, cognitive control and musical processing) make use of many of the same computations as high-level linguistic processing, even though they run them on duplicate hardware. Redundancy makes it is easy to see how fairly sharp dissociations could arise—knocking out one token module need

disrupt only one high-level operation: other high-level operations that draw on the same *type* of resource may well be spared.

The consequences of this distinction between literal reuse and reuse by proxy for much speculation about the localization and specialization of function are potentially profound. In cognitive neuropsychology the discovery that a focal lesion selectively impairs a particular cognitive function is routinely taken as evidence of its functional specificity (Coltheart 2011; Sternberg 2011). Even cognitive scientists who take a developmental approach to modularity, i.e. who concede that parts of the mind may be modular but stress that modularization is a developmental process, concede too much when they imply, as they frequently do, that modularization results in domain-specific modules (Karmiloff-Smith 1992; Prinz 2006; Barrett 2006; Cowie 2008; Guida et al. 2016). This is true in some sense, but not in anything like the standard sense, for redundancy envisages that developmental modules form a special class of neural networks, namely those which are *qualitatively* identical but *numerically* distinct. The appearance of modularization in development is thus fully compatible with deep domain interpenetration. In any event redundancy does not predict that all acquired skills will be modular. The evidence suggests that while some complex skills reside in at least partly dissociable circuitry, most complex skills are implemented in more typical neural networks, i.e. those consisting of literally shared parts.[8]

[8] This seems to be true regardless of whether the complex skills are innate or acquired.

4. What Else Might Redundancy Explain?

It is generally a good design feature of any system to have spare capacity. For instance, in engineered systems, "redundant parts can substitute for others that malfunction or fail, or augment output when demand for a particular output increases" (Whiteacre 2010, 14). The positive connection between robustness and redundancy in biological systems is also clear (Edelman and Gally 2001; Mason 2010; Whiteacre 2010; Iriki & Taoka 2012). So there are good reasons for evolution to have seen to it that our brains have spare capacity. But in the case of the brain and the cortex most especially, there are other reasons why redundancy would be an important design feature. It offers a solution to what Jungé and Dennett (2010, 278) called the "time-sharing" problem. It may also offer a solution to what I call the "encapsulation" problem.

The time-sharing problem arises when multiple simultaneous demands are made on the same cognitive resource. This is probably a regular occurrence. Here are just a few examples.

- Driving a car and holding a conversation at the same time: if it is true that some of the selfsame motor operations underlying aspects of speech production and comprehension are also required for the execution of sequenced or complex motor functions (Pulvermüller and Fadiga 2010; Graziano et al. 2002; MacNeilage 1998; Glenberg et al.

2008; Glenberg and Kaschak 2002; Glenberg et al. 2007; Greenfield 1991), as perhaps exemplified by driving a manual vehicle or operating complex machinery (e.g. playing the organ), how do we manage to pull this off?

- By reflecting the recursive structure of thought (Christiansen and Chater 2016, 51), the language circuits may redeploy a recursive operation simultaneously during sentence production. This might be the case during the formation of an embedded relative clause—the thought and its encoding may require parallel use of the same sequencing principle. Again, how do we manage this feat?
- If metarepresentational operations are involved in the internalization of conventional sound-meaning pairs, and also in the pragmatics and mindreading that carry on simultaneously during conversation, as argued by Suddendorf (2013), could this not simply be another instance of time-sharing? The example is contentious, but it still raises the question: how does our brain manage to do things like this?
- Christiansen and Chater's (2016) "Chunk and Pass" model of language processing envisages *multilevel* and *simultaneous* chunking procedures. As they put it, "the challenge of language acquisition is to learn a dazzling sequence of rapid processing operations" (2016, 116). What must the brain be like to allow for this dazzling display?

Explaining these phenomena is difficult. Indeed when dealing with clear (literal) instances of reuse, results from the interference paradigm show that processing bottlenecks are inevitable—true multi-tasking is impossible. Redundancy offers a natural explanation of how

the brain overcomes the time-sharing problem. It explains, in short, how we are able to walk and chew gum at the same time.

Redundancy might also offer a solution to what I have called the encapsulation problem. The neural networks that implement cognitive functions are not likely to be characterized by informational encapsulation if they share their nodes with networks implementing other cognitive functions. This is because in sharing their nodes with these other systems they will prima facie have access to the information stored and manipulated by those other systems (Anderson 2010, 300). If, then, overlapping brain networks must share information (Pessoa 2016, 23), it would be reasonable to suppose that central and peripheral systems do *not* overlap. For peripheral systems, which are paradigmatically fast and automatic, would not be able to process inputs as efficiently if there were a serious risk of central system override—i.e. of beliefs and other central information getting in the way of automatic processing. But we know from the neuroimaging literature that quite often the brain networks implementing central and peripheral functions *do* overlap. This is puzzling in light of the degree of cognitive impenetrability that certain sensory systems still seem to exhibit—limited though it may be. If it is plausible to suppose that the phenomenon calls for segregated circuitry, redundancy could feature in a solution to the puzzle, since it naturally explains how the brain can make parallel use of the same resources. Neuroimaging maps might well display what appear to be overlapping brain regions between two tasks (one involving central information, the other involving classically peripheral operations), but the

overlap would not exist—there would be distinct albeit adjacent or interdigitated and nearly identical circuits recruited in each case. Of course there may be other ways around the encapsulation problem that do not require segregated circuitry: the nature and extent of the overlap is presumably important. But clearly redundancy opens up some fascinating explanatory possibilities.

To the extent that acquired skills must overcome both the time-sharing problem as well as the encapsulation problem—for acquired competencies are often able to run autonomously of central processes—we might expect that their neural implementations incorporate redundant tissue. In concluding, let me illustrate this point by offering a gloss on a particular account of how skills and expertise are acquired during development elaborated by Guida et al. (2016) and Anderson (2014). The process involved is called "search" (Anderson 2014). Search is an exploratory synaptogenetic process, "the active testing of multiple neuronal combinations until finding the most appropriate one for a specific skill, i.e., the neural niche of that skill" (Guido et al. 2016, 13). The theory holds that in the early stages of skill acquisition, the brain must search for an appropriate mix of brain areas, and does so by recruiting relatively widely across the cortex. When expertise has finally developed, a much narrower and more specific network of brain areas has been settled upon, such that "[a]s a consequence of their extended practice, experts develop domain-specific knowledge structures" (Guido et al. 2016, 13). The gloss (and my hunch) is this: first, that repeated practice of a task that requires segregation (to get around time-

sharing and encapsulation issues) will in effect *force* search into redundant neural territory (Karmiloff-Smith 1992; Barrett 2006; Barret and Kurzban 2006); second, that search will recruit idle or relatively underutilized circuits in preference to busy ones as a general default strategy. Guido et al. (2016) cite evidence that experts' brains reuse areas for which novices' brains make only limited use: "Whereas novices use episodic long-term memory areas (e.g., the mediotemporal lobe) for performing long-term memory tasks, experts are able to (re)use these areas also for performing working-memory tasks" (Guido et al. 2016, 14). Guido and colleagues, in agreement with Anderson (2014), seem to have literal reuse in mind. But the same evidence they cite is consistent with reuse by proxy. As Barrett and Kurzban (2006, 639) suggest, echoing a similar suggestion by Karmiloff-Smith (1992), a developmental system

> could contain a procedure or mechanism that partitioned off certain tasks—shunting them into a dedicated developmental pathway—under certain conditions, for example, when the cue structure of repeated instances of the task clustered tightly together, and when it was encountered repeatedly, as when highly practiced....Under this scenario, reading could still be recruiting an evolved system for object recognition, and yet phenotypically there could be *distinct modules* for reading and for other types of object recognition.

5. Conclusion

It is true that language and other cognitive skills frequently dissociate from other skills, but redundancy puts this sort of modularization in its proper context. Redundancy predicates functional inheritance across tasks and task categories even when the tasks are implemented in spatially segregated neural networks. Thus dissociation evidence alone does not always indicate true functional specificity. In particular, these dissociations provide no evidence that language is cognitively special vis-à-vis other cognitive domains.

References

Anderson, Michael L. 2007a. "Evolution of Cognitive Function via Redeployment of Brain Areas." *The Neuroscientist* 13:13-21.

——2007b. "Massive Redeployment, Exaptation, and the Functional Integration of Cognitive Operations." *Synthese* 159 (3): 329-345.

——2007c. "The Massive Redeployment Hypothesis and the Functional Topography of the Brain." *Philosophical Psychology* 21 (2): 143-174.

——2008. "Circuit Sharing and the Implementation of Intelligent Systems." *Connection Science* 20 (4): 239-251.

——2010. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences* 33 (4): 245-266; discussion 266-313.

——2014. *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Anderson, Michael L., and Barbara L. Finlay. 2014. "Allocating Structure to Function: The Strong Links Between Neuroplasticity and Natural Selection." *Frontiers in Human Neuroscience* 7:1-16.

Apperly, I.A., D. Samson, N. Carroll, S. Hussain, and G. Humphreys. 2006. "Intact First- and Second-Order False Belief Reasoning in a Patient with Severely Impaired Grammar." *Social Neuroscience* 1 (3-4): 334-348.

Barrett, H. Clark. 2006. "Modularity and Design Reincarnation." In *The Innate Mind Volume 2: Culture and Cognition*, ed. Peter Carruthers, Stephen Laurence, and Stephen P. Stich, 199-217. New York: Oxford University Press.

Barrett, H. Clark, and Robert Kurzban. 2006. "Modularity in Cognition: Framing the Debate." *Psychological Review* 113 (3): 628-647.

Bergeron, Vincent. 2007. "Anatomical and Functional Modularity in Cognitive Science: Shifting the Focus." *Philosophical Psychology* 20 (2): 175-195.

Buxhoeveden, Daniel P., and Manuel F. Casanova. 2002. "The Minicolumn Hypothesis in Neuroscience." *Brain* 125:935-951.

Carruthers, Peter. 2006. The *Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Casasanto, D., and K. Dijkstra. 2010. "Motor Action and Emotional Memory." *Cognition* 115 (1): 179-185.

Chomsky, Noam. 1980. *Rules and Representations*. New York: Columbia University Press.

——1988. *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.

——2002. *On Nature and Language*. New York: Cambridge University Press.

Christiansen, Morten H., and Nick Chater. 2016. *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge, MA: MIT Press.

Coltheart, Max. 1999. "Modularity and Cognition." *Trends in Cognitive Sciences* 3 (3): 115-120.

——2011. "Methods for Modular Modelling: Additive Factors and Cognitive Neuropsychology." *Cognitive Neuropsychology* 28 (3-4): 224-240.

Cosmides, Leda, and John Tooby. 1994. "Origins of Domain Specificity: The Evolution of Functional Organization." In *Mapping the World: Domain Specificity in Cognition and Culture*, ed. L. Hirschfield, and S. Gelman, 85-116. New York: Cambridge University Press.

Cowie, Fiona. 2008. "Innateness and Language." In *The Stanford Encyclopedia of Philosophy*, winter 2016, ed. E.N. Zalta. <http://plato.stanford.edu/archives/win2016/entries/innateness-language/>

de Almeida, Roberto G., and Lila R. Gleitman, eds. 2018. *On Concepts, Modules, and Language: Cognitive Science at its Core*. New York: Oxford University Press.

Deacon, Terrence W. 2010. "A Role for Relaxed Selection in the Evolution of the Language Capacity." *Proceedings of the National Academy of Sciences of the United States of America* 107: 9000-9006.

Dehaene, Stanislas. 2005. "Evolution of Human Cortical Circuits for Reading and Arithmetic: The 'Neuronal Recycling' Hypothesis." In *From Monkey Brain to Human Brain*, eds. Stanislas Dehaene, J.R. Duhamel, M.D. Hauser, and G. Rizzolatti, 133-157. Cambridge, MA: MIT Press.

Edelman, Gerald M., and Joseph A. Gally. 2001. "Degeneracy and Complexity in Biological Systems." *Proceedings of the National Academy of Sciences of the United States of America* 98 (24): 13763-13768.

Eliasmith, Chris. 2015. "Building a Behaving Brain." In *The Future of the Brain*, ed. Gary Marcus, and Jeremy Freeman, 125-136. Princeton: Princeton University Press.

Fedorenko, Evelina, Michael K. Behr, and Nancy Kanwisher. 2011. "Functional Specificity for High-Level Linguistic Processing in the Human Brain." *Proceedings of the National Academy of Sciences of the United States of America* 108 (39): 16428-16433.

Fedorenko, Evelina, John Duncan, and Nancy Kanwisher. 2012. "Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area." *Current Biology* 22 (21): 2059-2062.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Friston, Karl J., and Cathy J. Price. 2003. "Degeneracy and Redundancy in Cognitive Anatomy." *Trends in Cognitive Sciences* 7 (4): 151-152.

Gauthier, I., T. Curran, K.M. Curby, and D. Collins. 2003. "Perceptual Interference Supports a Non-Modular Account of Face Processing." *Nature Neuroscience* 6 (4): 428-432.

Glenberg, A.M., M. Brown, and J.R. Levin. 2007. "Enhancing Comprehension in Small Reading Groups Using a Manipulation Strategy." *Contemporary Educational Psychology* 32:389-399.

Glenberg, A.M., and M.P. Kaschak. 2002. "Grounding Language in Action." *Psychonomic Bulletin and Review* 9:558-565.

Glenberg, A.M., M. Sato, and L. Cattaneo. 2008. "Use-Induced Motor Plasticity Affects the Processing of Abstract and Concrete Language." *Current Biology* 18 (7): R290-291.

Godfrey-Smith, Peter. 2001. "Three Kinds of Adaptationism." In *Adaptationism and Optimality*, ed. Steven H. Orzack, and Elliott Sober, 335-357. Cambridge: Cambridge University Press.

Graziano, M.S.A., C.S.R. Taylor, T. Moore, and D.F. Cooke. 2002. "The Cortical Control of Movement Revisited." *Neuron* 36:349-362.

Greenfield, P.M. 1991. "Language, Tools and Brain: The Ontogeny and Phylogeny of Hierarchically Organized Sequential Behavior." *Behavioral and Brain Sciences* 14 (4): 531- 551; discussion 551-595.

Guida, Alessandro, Guillermo Campitelli, and Fernand Gobet. 2016. "Becoming an Expert: Ontogeny of Expertise as an Example of Neural Reuse." *Behavioral and Brain Sciences* 39:13-15.

Hickok, G., and David Poeppel. 2000. "Towards a functional neuroanatomy of speech perception." *Trends in Cognitive Sciences* 4 (4): 131-138.

Iriki, Atsushi, and Miki Taoka. 2012. "Triadic (ecological, neural, cognitive) niche construction: A scenario of human brain evolution extrapolating tool use and language from the control of reaching actions." *Philosophical Transactions of the Royal Society B* 367: 10-23.

Jungé, Justin A., and Daniel C. Dennett. 2010. "Multi-Use and Constraints from Original Use." *Behavioral and Brain Sciences* 33 (4): 277-278.

Kandel, E.R., and A.J. Hudspeth. 2013. "The Brain and Behavior." In *Principles of Neural Science*, ed. E.R. Kandel, J.H. Schwartz, T.M. Jessell, S.A. Siegelbaum, and A.J. Hudspeth, 5-20. New York: McGraw-Hill.

Karmiloff-Smith, Annette. 1992. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Leo, Andrea, Giulio Bernardi, Giacomo Handjaras, Daniela Bonino, Emiliano Ricciardi, and Pietro Pietrini. 2012. "Increased BOLD Variability in the Parietal Cortex and Enhanced Parieto-Occipital Connectivity During Tactile Perception in Congenitally Blind Individuals." *Neural Plasticity* 2012:1-8 doi: 10.1155/2012/720278.

Luria, A.R., L.S. Tsvetkova, and D.S. Futer. 1965. "Aphasia in a Composer (V.G. Shebalin)." *Journal of the Neurological Sciences* 2 (3): 288-292.

MacNeilage, P.F. 1998. "The Frame/Content Theory of Evolution of Speech Production." *Behavioral and Brain Sciences* 21 (4): 499-511; discussion 511-546.

Maleszka, Ryszard, Paul H. Mason, and Andrew B. Barron. 2013. "Epigenomics and the Concept of Degeneracy in Biological Systems." *Briefings in Functional Genomics* 13 (3): 191-202.

Mason, Paul H. 2010. "Degeneracy at Multiple Levels of Complexity." *Biological Theory* 5 (3): 277-288.

McGeer, Victoria. 2007. "Why Neuroscience Matters to Cognitive Neuropsychology." *Synthese* 159:347-371.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Pessoa, Luiz. 2016. "Beyond Disjoint Brain Networks: Overlapping Networks for Cognition and Emotion." *Behavioral and Brain Sciences* 39:22-24.

Peretz, Isabelle., and Max Coltheart. 2003. "Modularity of music processing." *Nature Neuroscience* 6:688-691.

Pinker, Steven, and Ray Jackendoff. 2005. "The Faculty of Language: What's Special About It?" *Cognition* 95:201-236.

Plaut, David C. 1995. "Double Dissociation Without Modularity: Evidence from Connectionist Neuropsychology." *Journal of Clinical and Experimental Psychology* 17 (2): 291-321.

Poeppel, David. 2001. "Pure Word Deafness and the Bilateral Processing of the Speech Code." *Cognitive Science* 21 (5): 679-693.

Prinz, Jesse J. 2006. "Is the Mind Really Modular?" In *Contemporary Debates in Cognitive Science*, ed. R. Stainton, 22-36. Oxford: Blackwell.

Pulvermüller, Friedmann, and Luciano Fadiga. 2010. "Active Perception: Sensorimotor Circuits as a Cortical Basis for Language." N*ature Reviews Neuroscience* 11:351-360.

Sternberg, Saul. 2011. "Modular Processes in Mind and Brain." *Cognitive Neuropsychology* 28 (3-4): 156-208.

Striem-Amit, Ella, and Amir Amedi. 2014. "Visual Cortex Extrastriate Body-Selective Area Activation in Congenitally Blind People 'Seeing' by Using Sounds." *Current Biology* 24:1-6.

Suddendorf, Thomas. 2013. *The Gap: The Science of What Separates Us from the Animals*. New York: Basic Books.

Varley, R.A., N.J.C. Klessinger, C.A.J. Romanowski, and M. Siegal. 2005. "Agrammatic But Numerate." *Proceedings of the National Academy of Sciences of the United States of America* 102:3519-3524.

Whiteacre, James M. 2010. "Degeneracy: A Link Between Evolvability, Robustness and Complexity in Biological Systems." *Theoretical Biology and Medical Modelling* 7 (6): 1-17.

Zerilli, John. 2017a. "Against the 'System' Module." *Philosophical Psychology* 30 (3): 235-250.

——2017b. "Multiple Realization and the Commensurability of Taxonomies." *Synthese* (https://doi.org/10.1007/s11229-017-1599-1).

Supervenience, reduction, and translation

February 23, 2018

This paper considers the following question: what is the relationship between supervenience and reduction? I investigate this formally, first by introducing a recent argument by Christian List to the effect that one can have supervenience without reduction; then by considering how the notion of Nagelian reduction can be related to the formal apparatus of definability and translation theory; then by showing how, in the context of propositional theories, topological constraints on supervenience serve to enforce reducibility; and finally, how constraints derived from the theory of ultraproducts can enforce reducibility in the context of first-order theories.

1 List on supervenience and reduction

I'll start by giving a brief recapitulation of the apparatus used by List (2018) to analyse supervenience. For List, a *system of ontological levels* is a concrete posetal category wherein each function is surjective. That is, it consists of a class of sets $\mathcal{L}$, equipped with surjective functions between them, such that (i) the class of all such functions is closed under composition, and for each set we include the identity function on that set (hence making this a concrete category); and (ii) for any sets A and B, there is a *unique* function σ mapping A to B (hence making the category posetal).

In order to relate this to issues of reduction, List introduces certain linguistic notions. Formally, he defines a *language* L to be a set (of "sentences"), equipped with a unary "negation" operator and a "consistency" property. Given a language L, List defines the *ontology* Ω_L of L to be the set of all maximal consistent subsets of L (where a consistent set A is *maximal* iff there is no consistent B such that $A \subset B$). Each element of this ontology he calls a *world* for L. For any $\phi \in L$, ϕ is *true at* the world $\omega \in \Omega_L$ iff $\phi \in \omega$; otherwise, it is false at ω. The *propositional content* $[[\phi]]$ of $\phi \in L$ is the set of worlds at

which ϕ is true: that is, $[[\phi]] := \{\omega \in \Omega_L : \phi \in \omega\}$. Then, a *system of descriptive levels* is a system of ontological levels where each level is the ontology Ω_L of some language L.

The notion of reduction is defined as follows. Given some system of descriptive levels, suppose that $\sigma : \Omega_L \to \Omega_{L'}$ is a supervenience mapping. List says that a sentence $\phi' \in L'$ *reduces* to $\phi \in L$ if $[[\phi]] = \sigma^{-1}([[\phi']])$: that is, if ϕ is true at a world ω iff ϕ' is true at $\sigma(\omega)$. And we say that the language L' reduces to L if every $\phi' \in L$ reduces to some $\phi \in L$: that is, if every higher-level sentences is reducible to some lower-level sentence.

Certainly, setting up this kind of association between the apparatus of possible worlds and languages is going to be necessary if we are going to talk about the relationship between supervenience (a map, on this treatment, between ontologies) and reduction (a relationship between languages). But it is not clear to me that this is quite the right way to set things up.

First, it seems a mistake to use maximal consistent sets of sentences to represent the ontology associated to a language. This is not because I have a problem with identifying worlds with sets of sentences—we're just doing mathematical modelling, after all, so "identification" isn't really doing anything more than asserting a one-to-one correspondence. Rather, the problem is that there are intuitive reasons to think that, in general, there are more worlds (of a given language's ontology) than there are maximal consistent sets of sentences of that language. Specifically, consider the class of models of a first-order language L. It seems plausible to suppose that there could be a world just like each model—one with the same number of individuals as there are elements in the model's domain, and with properties and relations distributed over those individuals just as the extensions of the predicates are distributed over the model. This suggests that we can take such models to represent worlds, with non-isomorphic models representing distinct worlds.[1] But in first-order logic, there are non-isomorphic models which are elementarily equivalent, i.e., which satisfy all the same sentences. So a maximal consistent set of sentences will, in general, be satisfied by several non-isomorphic models, and so by several distinct worlds.

For this reason, I suggest that a friendly amendment to List's proposal is that we take the ontology associated with a language to be the class of models for that language, rather than the set of maximal consistent sets of sentences of that language. This will require us to say a little more about the nature of the language: for the purposes of this essay, I will assume that we are working with finitary first-order languages, each

[1] Whether isomorphic models represent distinct worlds or not is more controversial (it is closely related to the question of whether or not to believe in haecceities, i.e., primitive trans-world identities for individuals).

with a particular nonlogical vocabulary. Note that we are including the possibility that the language is a propositional language (regarding propositional constants as nullary predicates).

That said, it does seem that we don't want to represent the ontology associated to a given level by *all* models of the associated language. After all, if the language contains predicates R for "red all over" and G for "green all over", then there will be a model containing objects satisfying both R and G. That is, the worlds represented by the class of all models is intuitively the class of logically possible worlds, but that is typically not the species of possibility that we are interested in for the purposes of supervenience. Standardly, we want to consider supervenience relative to metaphysical possibility, or relative to something more restrictive still (e.g. physical possibility). In List's approach, one could code this up into the consistency property, so that only worlds corresponding to metaphysically or physically consistent sets of sentences are permitted (where "physical inconsistency" would mean, roughly, consistency with the physical laws).

In the friendly amendment I'm suggesting here, a more natural way to do things would be to confine attention to the models *of some theory*, where that theory consists of all and only the sentences expressing things which are necessary—necessary, that is, relative to the standard of possibility we seek to capture. Thus, if we are thinking about metaphysical possibility, then we take the theory to consist of all the metaphysically necessary L-sentences (e.g. $\forall x(\neg Rx \wedge Gx)$); if we are thinking about physical possibility, then we take the theory to consist of all the L-sentences expressing physical laws; and so on and so forth. So every descriptive level is associated with a given theory, and the supervenience maps between levels are maps between the sets of models of the associated theories. We follow List in imposing the requirements that the sets of models, equipped with the supervenience maps, constitute a posetal category. Rather than requiring that these maps be surjective, however, it will be more helpful to reflect further on the relationship between models and the possible worlds they represent.

The relevant observation is that consistency with the laws at its own level is only a necessary condition for a model to represent a possible world, not a sufficient one: it is natural to think that the sense of possibility relevant for supervenience is something stronger. Just because a certain distribution of thermodynamical properties is possible according to the laws of thermodynamics, for instance, does not mean that that distribution is possible *tout court*; in particular, such a distribution may not be one that could actually be generated by any statistical-mechanically possible configuration of the microstructure. Thus, in the event that two descriptive levels, associated with theories

T_1 and T_2, have a supervenience map $\sigma : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$, then we should take the ontology of T_1 (i.e., the possible worlds at the T_1-level) to include only those models of T_1 that lie in the image of the supervenience mapping: i.e., that $\Omega_1 \subseteq \sigma[\mathrm{Mod}(T_2)]$. And of course, for any other descriptive level upon which the T_1-level supervenes, the same requirement will hold.

By taking these to be the only such requirements, we may define the ontology associated with a given level as follows: suppose that the T-level supervenes upon all and only the T_i-levels, for i in some index set I, with the supervenience mappings $\sigma_i : \mathrm{Mod}(T_i) \to \mathrm{Mod}(T)$. Then we let the ontology for the T-level be $\Omega := \bigcap_{i \in I} \sigma_i[\mathrm{Mod}(T_i)]$. It follows that if the T_1-level supervenes upon the T_2-level, then the supervenience map is surjective on ontologies: that is, for every $\mathcal{M}_1 \in \Omega_1$, there is some $\mathcal{M}_2 \in \Omega_2$ such that $\sigma(\mathcal{M}_2) = \mathcal{M}_1$.

Finally, if the T_1-level supervenes upon the T_2-level, with supervenience map σ, I'll say that an L_1-sentence ϕ_1 *List-reduces* to the L_2-sentence ϕ_2 if $[[\phi_2]] = \sigma^{-1}([[\phi_1]])$: that is, if for every $\mathcal{M} \in \Omega_2$, $\mathcal{M} \models \phi_2$ iff $\sigma(\mathcal{M}) \models \phi_1$. And T_1 *List-reduces* to T_2 if every L_1-sentence ϕ_1 is reducible to some L_2-sentence ϕ_2.

List observes that, within his framework, supervenience does not entail reducibility. In the version of that framework developed here, this is encoded by the following result.

Theorem 1. *Not every system of descriptive levels is such that if the T_1-level supervenes on the T_2-level, then T_1 is List-reducible to T_2.*

List appeals to general combinatorial considerations for a proof. However, it will be illuminating (and helpful for the sequel) to construct an explicit counterexample to the hypothesis that supervenience entails reduction.[2]

Proof. Let L_1 be the propositional language whose only sentence-letter is F, and let L_2 be the propositional language with sentence-letters $\{P_0, P_1, \dots\}$. Let $T_1 = T_2 = \varnothing$. $\mathrm{Mod}(T_1)$ only contains two worlds: set $\mathcal{M}_F(F) = \top$, and $\mathcal{M}_{\neg F}(F) = \bot$. Let $\mathcal{M}$ be the T_2-model such that $\mathcal{M}(P_i) = \top$ for every P_i. Define $\sigma : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ as follows: for any $\mathcal{A} \in \mathrm{Mod}(T_2)$,

$$\sigma(\mathcal{A}) := \begin{cases} \mathcal{M}_F & \text{if } \mathcal{A} = \mathcal{M} \\ \mathcal{M}_{\neg F} & \text{otherwise} \end{cases} \tag{1}$$

Since the T_2-level is the lowest, $\Omega_2 = \mathrm{Mod}(T_2)$; hence $\Omega_1 = \sigma[\Omega_2] = \mathrm{Mod}(T_1)$.

[2]This example is inspired by a construction in Halvorson (2012).

Now observe that $\sigma^{-1}([[F]]) = \{\mathcal{M}\}$. But $\{\mathcal{M}\}$ is not a definable subset of Ω_2: that is, there is no sentence $\phi \in L_2$ such that $[[\phi]] = \{\mathcal{M}\}$.[3] So the sentence $F \in L_1$ is not reducible to any sentence in L_2; thus, T_1 is not reducible to T_2. □

Now, one natural (and I think appropriate) response to this argument is that it is unsurprising. After all, we have imposed almost no constraints on the supervenience map. A supervenience map between two ontological levels is just a map from the worlds of one level to the worlds of the other, with no constraints on how that map must relate to the structure present in those worlds. So for the rest of this essay, I would like to consider some ways in which we could impose such constraints on the maps, and what consequences those constraints have for reducibility.

2 Definition and reduction

Before going further, I want to pause to make contact with some more fully-fledged ways of thinking about reduction. This will involve some of the apparatus of definability theory from logic.[4] An *explicit definition* of R in terms of Σ is a formula of the form

$$\delta_R = \forall x_1 \dots \forall x_n (Rx_1 \dots x_n \leftrightarrow \tau_R(x_1, \dots, x_n)) \tag{2}$$

where τ_R is a Σ-formula. Given a Σ-theory T and Σ^+-theory T^+, where $\Sigma \subset \Sigma^+$, T^+ is a *definitional extension* of T if T^+ is logically equivalent to the theory

$$T \cup \{\delta_R : R \in \Sigma^+ \setminus \Sigma\} \tag{3}$$

where for each R, δ_R is an explicit definition of R in terms of Σ. More generally, any Σ^+-theory T^+ explicitly defines a symbol $R \in \Sigma^+ \setminus \Sigma$ in terms of Σ if it entails an explicit definition δ_R of R in terms of Σ: that is, if $T^+ \vDash \delta_R$ for some δ_R of the form (2).

Now, this may be compared to Nagel (1979)'s definition of reduction:

> [...] a reduction is effected when the experimental laws of the secondary science (and if it has an adequate theory, its theory as well) are shown to be the logical consequences of the theoretical assumptions (inclusive of the

[3] Proof: suppose for reductio that $[[\phi]] = \{\mathcal{M}\}$. Since ϕ is a finite sentence, not every sentence-letter can occur in it. So suppose P_i does not occur in ϕ. Then since $\mathcal{M} \models \phi$, it must be the case that $\mathcal{M}' \models \phi$, where $\mathcal{M}'$ is just like $\mathcal{M}$ save that $\mathcal{M}'(P_i) = \bot$ (as the truth-value of a sentence in propositional logic is dependent only on the truth-values of the sentence-letters occurring in it). But then $\mathcal{M}' \in [[\phi]]$, although $\mathcal{M}' \neq \mathcal{M}$, so we have a contradiction. QED.

[4] See (Hodges, 1997, §2.6) for further details.

> coordinating definitions) of the primary science. [. . .] when the laws of the secondary science do contain some term 'A' that is absent from the theoretical assumptions of the primary science, there are two necessary formal conditions for the reduction of the former to the latter: (1) Assumptions of some kind must be introduced which postulate suitable relations between whatever is signified by 'A' and traits represented by terms already present in the primary science. [. . .] (2) With the help of these additional assumptions, all the laws of the secondary science, including those containing the term 'A', must be logically derivable from the theoretical premises and their associated coordinating definitions in the primary discipline.[5]

As it is often put: a theory T_1 is reducible to a theory T_2 just in case T_1 can be derived from T_2 together with so-called *bridge laws*. Now, as has been widely discussed,[6] if we place no restrictions on the content of the bridge laws then reduction becomes trivialised: any theory T_1 may be reduced to any other theory T_2, by taking as the bridge laws either T_1 itself, or any inconsistent set of statements. To avoid this problem, we put two restrictions on the bridge laws. The first is that the bridge laws do indeed serve to bridge the gap between T_1 and T_2: formally, we require that T_2 plus the bridge laws explicitly defines every symbol $R \in \Sigma_1 \setminus \Sigma_2$ (where Σ_1 and Σ_2 are, respectively, the vocabularies of T_1 and T_2). The second is that the bridge laws not sneak in "extra content" beyond this bridging function: formally, we require that T_2 plus the bridge laws be a *conservative extension* of T_2. (Recall that a Σ^+-theory T^+ is a conservative extension of a given Σ-theory T, where $\Sigma \subseteq \Sigma^+$, if T^+ has no new Σ-consequences relative to T: that is, if for every Σ-sentence ϕ, $T^+ \models \phi$ entails that $T \models \phi$.)

But if these two conditions are satisfied, then T_2 together with the bridge laws is a definitional extension of T_2.[7] Thus, to within logical equivalence, we may suppose that the bridge laws *are* explicit definitions of the novel terms (i.e., of the symbols in $\Sigma_1 \setminus \Sigma_2$).[8] Note that this agrees with Schaffner's requirement that bridge laws be "reduction functions" (statements that a certain term of T_1 and a certain term of T_2 are coextensional)—provided, that is, that we take the reduction functions to relate a simple term of T_1 to a (possibly) complex term of T_2.[9] Thus, we are led to posit the following

[5](Nagel, 1979, pp. 352–354)

[6]See, for instance, Dizadji-Bahmani et al. (2010) and references therein.

[7](Hodges, 1997, pp. 53–54).

[8]Note that all we are supposing here is that every bridge law has the syntactic form of a definition (i.e., of a biconditional of the form (2)). This is neutral on the question of what the *content* of the bridge laws is: for instance, whether that of a set of definitions, a set of conventional stipulations, or a set of factual assertions (Nagel, 1979, pp. 354–355).

[9]This identification of Nagelian reduction with definitional extension is also made by Butterfield (2011a).

relation: a theory T_1 *Nagel-reduces* to T_2 if there is some definitional extension T_2^+ of T_2 such that $T_2^+ \vDash T_1$.[10]

We now want to think about how all this relates to the apparatus of supervenience mappings developed in the previous section. For these purposes, it is helpful to work with the notions of interpretation and translation rather than definition—though as we shall see, these are closely intertwined.[11] So, define an *interpretation* from one language L_1 to another language L_2 to be a map $\tau : \mathrm{Form}(L_1) \to \mathrm{Form}(L_2)$ which commutes with the logical constants (where $\mathrm{Form}(L)$ is the set of formulae of L). And given theories T_1 and T_2, in languages L_1 and L_2 respectively, define a *translation* from T_1 to T_2 to be an interpretation $\tau : L_1 \to L_2$ such that for any $\phi \in \mathrm{Form}(L_1)$, if $T_1 \vDash \phi$, then $T_2 \vDash \tau(\phi)$. This lets us give a more compact characterisation of Nagel-reduction, as follows.

Theorem 2. *T_1 Nagel-reduces to T_2 iff there is a translation $\tau : T_1 \to T_2$ which restricts to the identity on $\Sigma_1 \cap \Sigma_2$.*

Proof. Before starting the proof proper, we state a quick lemma. Suppose that $\tau : \Sigma_1 \to \Sigma_2$ is an interpretation which restricts to the identity on $\Sigma_1 \cap \Sigma_2$. For any n-ary $R \in \Sigma_1 \setminus \Sigma_2$, let $\delta_R := \forall x_1 \ldots \forall x_n (Rx_1 \ldots x_n \leftrightarrow \tau(R)(x_1, \ldots, x_n))$. Then for any $\phi \in \mathrm{Form}(\Sigma_1 \cup \Sigma_2)$, $\{\delta_R : R \in \Sigma_1 \setminus \Sigma_2\} \vDash \phi \leftrightarrow \tau(\phi)$. The lemma can be demonstrated by a straightforward proof by induction on the complexity of formulae.

Now, suppose that T_1 Nagel-reduces to T_2, i.e., that $T_2^+ \vDash T_1$ for some definitional extension T_2^+ of T_2. Then for any $R \in \Sigma_1 \setminus \Sigma_2$, $T_2^+ \vDash \delta_R$ where δ_R is some explicit definition of R, i.e., some formula of the form (2). So let $\tau(R) := \tau_R$ for any such R (where τ_R is the Σ_2-formula occurring on the right-hand-side of the biconditional in δ_R). For any $S \in \Sigma_1 \cap \Sigma_2$, let $\tau(S) = S$. Now extend τ to a map from $\mathrm{Form}(\Sigma_1)$ to $\mathrm{Form}(\Sigma_2)$ by demanding that it commute with the logical constants. So τ is an interpretation from Σ_1 to Σ_2 (which restricts to the identity on $\Sigma_1 \cap \Sigma_2$), and it only remains to show that it is a translation from T_1 to T_2.

So for an arbitrary Σ_1-formula ϕ, suppose $T_1 \vDash \phi$. Then $T_2^+ \vDash \phi$, since $T_2^+ \vDash T_1$. By the lemma, $T_2^+ \vDash \phi \leftrightarrow \tau(\phi)$, and hence $T_2^+ \vDash \tau(\phi)$. But $\tau(\phi)$ is a Σ_2-formula, and T_2^+ is a conservative extension of T_2; so $T_2 \vDash \tau(\phi)$. This suffices to show that τ is a translation, as requested.

Next, suppose that τ is a translation from T_1 to T_2 which restricts to the identity on $\Sigma_1 \cap \Sigma_2$. For every $R \in \Sigma_1 \setminus \Sigma_2$, let $\delta_R := \forall x_1 \ldots \forall x_n (Rx_1 \ldots x_n \leftrightarrow \tau(R)(x_1, \ldots, x_n))$.

[10] Note that we are only permitting the definition of new relations, and not of any new objects in the domain: in more metaphysical terms, we are considering only reductions of properties and not of individuals. Obviously, this is rather limiting (the individuals of physics are clearly distinct from those of biology!)—but we will keep this restriction, in the interests of simplicity.

[11] cf. Barrett and Halvorson (2016)

Now let $T_2^+ := T_2 \cup \{\delta_R : R \in \Sigma_1 \setminus \Sigma_2\}$. So clearly, T_2^+ is a definitional extension of T_2, and it remains only to show that $T_2^+ \vDash T_1$.

So consider any $\phi \in T_1$. Then $T_2 \vDash \tau(\phi)$, since τ is a translation. So $T_2^+ \vDash \tau(\phi)$. By the lemma, $T_2^+ \vDash \phi \leftrightarrow \tau(\phi)$; so, $T_2^+ \vDash \phi$. This suffices to prove that $T_2^+ \vDash T_1$, and hence that T_1 Nagel-reduces to T_2. □

Let the class of Σ-structures be denoted $\mathrm{Str}(\Sigma)$. Now, given an interpretation $\tau : \Sigma_1 \to \Sigma_2$, the *dual map* of τ is a map $\tau^* : \mathrm{Str}(\Sigma_2) \to \mathrm{Str}(\Sigma_1)$, defined as follows: for any Σ_2-structure $\mathcal{A}$, $|\tau^*(\mathcal{A})| = |\mathcal{A}|$ (where $|\mathcal{M}|$ denotes the domain of the model $\mathcal{M}$); and for any $R \in \Sigma_1$ and $a_1, \ldots, a_n \in |\mathcal{A}|$, its extension $R^{\tau^*(\mathcal{A})}$ is the set of all and only those n-tuples which satisfy $\tau(R)$. It is straightforward to show that for any Σ_2-structure $\mathcal{A}$ and Σ_1-sentence ϕ, $\tau^*(\mathcal{A}) \models \phi$ iff $\mathcal{A} \models \tau(\phi)$. As a corollary, if τ is a translation from T_1 to T_2, then $\tau^* : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ (i.e., if $\mathcal{A}$ is a model of T_2, then $\mathcal{A}$ is a model of T_1).

Putting this all together, we get the following: if T_1 Nagel-reduces to T_2, then the reduction naturally induces a map from $\mathrm{Mod}(T_2)$ to $\mathrm{Mod}(T_1)$. Moreover, if the translations between a collection of theories constitute a posetal category, then their dual maps will as well: this follows from the fact that given translations $\tau_1 : T_1 \to T_2$ and $\tau_2 : T_2 \to T_3$, $(\tau_2 \circ \tau_1)^* = \tau_1^* \circ \tau_2^*$. So such a collection of translations gives rise to a system of descriptive levels of the kind motivated in the previous section. Thus, Nagel-reduction entails supervenience. Moreover, it entails List-reduction: given any L_1-sentence ϕ_1 and a translation $\tau : T_1 \to T_2$, ϕ_1 List-reduces to $\tau(\phi_2)$. It follows that the example in Theorem 1 also shows that supervenience does not entail Nagel-reducibility.

Incidentally, the dual maps will not typically be surjective (as maps between classes of models)—nor would it be appropriate to require that they were. To see this, observe that the following principle seems plausible: if T_1 reduces to T_2, and if T_0 is a subtheory of T_1 (i.e. if every sentence of T_0 is a sentence of T_1) then T_0 reduces to T_2. But it is an immediate consequence of this that not all dual maps can be surjective. Let T_0 be a proper subtheory of T_1, let τ be the translation from T_1 to T_2, and observe that it is thereby a translation from T_0 to T_2. If τ^* is surjective as a map from $\mathrm{Mod}(T_2)$ to $\mathrm{Mod}(T_1)$, then we are done. If not, then since T_0 is a proper subtheory of T_1, there is at least one model $\mathcal{M}$ of T_0 which is not a model of T_1. Since the image of $\mathrm{Mod}(T_2)$ under τ^* is $\mathrm{Mod}(T_1)$, then $\mathcal{M}$ is *not* in said image of, and so τ^* is not surjective as a map from $\mathrm{Mod}(T_2)$ to $\mathrm{Mod}(T_0)$.[12] As discussed previously, however, we can impose surjectivity of supervenience maps *as maps on possible worlds* by fiat.

[12] This poses a problem for Suppes' account of reduction (Suppes, 1957, 1967), given that he effectively presupposes surjectivity of the map from the lower-level theory to the higher-level theory associated with reduction.

With all this in hand, we now return to our main task: what kinds of constraints could be imposed on supervenience maps, in order to guarantee that they are associated with some form of reducibility? First, I will consider the idea that the supervenience maps must preserve *similarity*, as encoded in topological data about the models.

3 Supervenience and continuity

In his (2011b), Butterfield asks us to consider the sequence of functions $g_n : \mathbb{R} \to \mathbb{R}$, for $n \in \mathbb{N}^+ := \mathbb{N} \setminus \{0\}$, where

$$g_n(x) := \begin{cases} -1 & \text{if } x < -\frac{1}{n} \\ nx & \text{if } -\frac{1}{n} \leq x \leq \frac{1}{n} \\ 1 & \text{if } \frac{1}{n} < x \end{cases} \tag{4}$$

The limit of this sequence (as determined by pointwise convergence) is the function g_∞, where

$$g_\infty(x) := \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } 0 < x \end{cases} \tag{5}$$

As Butterfield notes, for each finite n the function g_n is continuous; yet the limit of the sequence of these individually continuous functions, g_∞, is discontinuous (at 0). Now consider the function $f_\bullet : \mathbb{N}^+ \cup \{\infty\} \to \{0, 1\}$, defined by

$$f_n := \begin{cases} 0 & \text{if } n \in \mathbb{N}^+ \\ 1 & \text{if } n = \infty \end{cases} \tag{6}$$

We can now define a two-level ontological system, in the sense of List, by taking one ontology to be the set $\{g_n\}_{n\in\mathbb{N}^+} \cup \{g_\infty\}$, the other to be the set $\{f_n\}_{n\in\mathbb{N}^+\cup\{\infty\}} = \{0, 1\}$, and the supervenience map to be $\sigma : g_n \mapsto f_n$. Alternatively, we could take the lower-level ontology to consist of all functions $\mathbb{R} \to \mathbb{R}$, the higher-level ontology to consist (again) of the set $\{0, 1\}$, and the supervenience function to be a function mapping continuous functions to 0 and discontinuous functions to 1.

This supervenience map fails to preserve limits. The limit of the sequence of lower-level worlds $g_1, g_2, \ldots$ is g_∞, and $\sigma(g_\infty) = 1$. By contrast, the limit of the sequence $\sigma(g_1), \sigma(g_2), \ldots$ is the limit of the sequence $0, 0, \ldots$, which is of course 0. So it makes a

difference whether we take the limit before or after applying the supervenience map; taking the limit does not commute with supervenience. Thinking of topological structure as giving a standard of similarity among the models,[13] let us say that a supervenience map which fails to commute with limits in this fashion is not *similarity-preserving*.

But now consider again the counterexample constructed in the proof of Theorem 1. We can make this example resembles Butterfield's example of the functions, by defining the models $\mathcal{M}_n$, for $n \in \mathbb{N}$, as follows: $\mathcal{M}_n$ assigns $\top$ to P_0 through P_n, and $\bot$ to all remaining sentence-letters. Then there is an intuitive sense in which the models in the sequence $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots$ get "more and more similar" to the model $\mathcal{M}$—that is, a sense in which we can regard $\mathcal{M}$ as the limit of the sequence $\mathcal{M}_n$ as n goes to infinity. This intuition receives a precise expression in the concept of the *Stone topology* of a class of propositional models: in that topology, a sequence of models $\mathcal{M}_n$ converges to a model $\mathcal{M}$ if and only if the sequence of truth-values $P^{\mathcal{M}_n}$ converges to the truth-value $P^{\mathcal{M}}$, for every sentence-letter P.[14] However, the supervenience map does not play nice with this convergence structure: just as with Butterfield's example, the supervenient image of the limit of the sequence (i.e., $\mathcal{M}_F$) is not the limit of the supervenient images of the sequence (i.e., $\mathcal{M}_{\neg F}$). In other words, the supervenience map is not similarity-preserving.

This suggests a conjecture: perhaps all such failures of reduction are associated with the failure of the supervenience map to be similarity-preserving? In the propositional case, this turns out to be correct. More specifically, we have the following result.

Theorem 3. *Suppose that T_1 and T_2 are propositional theories. Then there is a translation $\tau : T_1 \to T_2$ iff there is a map $\sigma : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ which is continuous in the Stone topology. In the event that this holds, $\sigma = \tau^*$.*

Proof. First, suppose that τ is a translation from T_1 to T_2. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the Lindenbaum-Tarski algebras of T_1 and T_2.[15] By Stone's representation theorem, any homomorphism $\rho : \mathcal{T}_1 \to \mathcal{T}_2$ is associated with a (Stone-)continuous function $\sigma : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ and vice versa. So in one direction, the translation $\tau : T_1 \to T_2$ gives rise to a homomorphism $\rho : \mathcal{T}_1 \to \mathcal{T}_2$, and thence to a continuous function σ, which coincides with the dual map τ^*. And in the other, a continuous function σ gives rise to a homomorphism ρ; but any homomorphism between atomic Boolean algebras

[13] cf. Fletcher (2016)

[14] Halvorson (2012)

[15] The Lindenbaum-Tarski algebra of a propositional theory T is a Boolean algebra whose elements are T-equivalent sets of sentences, and whose meet, join and complement are given by (the abstractions of) the conjunction, disjunction and negation operators.

is generated by a map from the atoms of one to those of the other, and so ρ is generated by a translation $\tau : T_1 \to T_2$, which is such that $\tau^* = \sigma$. □

4 The first-order case

Let us now turn our attention to the case of first-order theories. It will be helpful to split our enquiry into two parts. First, can we find conditions on a supervenience mapping sufficient for that mapping to be associated with a reduction? Second, can we think of those conditions as imposing a requirement of similarity-preservation?

With regard to the first question, the following result—based on a result by van Benthem and Pearce (1984)—provides an affirmative answer.

Theorem 4. *Suppose that T_1 and T_2 are first-order theories. Then there is a translation $\tau : T_1 \to T_2$ iff there is a map $\sigma : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ which preserves isomorphisms, ultraproducts and domains.*[16] *In the event that this holds, $\sigma = \tau^*$.*

Before turning to the proof, I review the relevant notion of an ultraproduct.[17] First, an *ultrafilter* over a set X is a non-empty set U of subsets of X, such that

- $\varnothing \notin U$
- U is closed under intersection: if $A \in U$ and $B \in U$, then $A \cap B \in U$
- U is closed under supersets: if $A \in U$ and $A \subseteq B$ (for $B \subseteq X$), then $B \in U$
- For every subset A of X, exactly one of A and $X \setminus A$ is in U.

Roughly speaking, one can think of an ultrafilter as equipping X with a notion of "almost all"-ness: relative to U, a subset $A \subseteq X$ contains "almost all" elements of X just in case $A \in U$.

Given an ultrafilter U on some set I (which we will call the set of indices), and an indexed family $\{\mathcal{M}_i : i \in I\}$ of Σ-structures, the ultraproduct $\Pi_U \mathcal{M}_i$ of the family is a Σ-structure defined as follows. First, form the Cartesian product $\Pi_{i \in I} |\mathcal{M}_i|$ of the domains of the structures in the family: that is, each element of $\Pi_{i \in I} |\mathcal{M}_i|$ is an indexed family of elements $\mathbf{a} = \{a^i \in |\mathcal{M}_i| : i \in I\}$, one from each $\mathcal{M}_i$. Then define the following equivalence relation: $\mathbf{a} \sim \mathbf{b}$ iff the set of indices i such that $a^i = b^i$ is in the ultrafilter. That is, $\mathbf{a}$ and $\mathbf{b}$ are equivalent if they are identical at "almost all" models in the family of

[16] As discussed above (see footnote 10), the condition of domain-preservation is clearly not very appropriate to the context of scientific theories—but is helpful in simplifying the technicalities.

[17] For an introduction to ultraproducts, see (Hodges, 1997, §8.5).

structures. We then take the domain of the ultraproduct to be the quotient of $\Pi_{i\in I}|\mathcal{M}_i|$ by this equivalence relation (so its elements are sets of "almost-everywhere-identical" indexed families). Finally, we determine the extension of any Σ-symbol R as follows: for n-ary R, the tuple $\langle \mathbf{a}_1, \ldots, \mathbf{a}_n\rangle$ is in the extension of R if and only if the set of indices i such that $\langle a_1^i, \ldots, a_n^i\rangle \in R^{\mathcal{M}_i}$ is in U. Thus, for instance, an indexed family $\mathbf{a}$ satisfies a unary predicate iff it satisfies the predicate at "almost all" models in the family of structures. An ultraproduct all of whose factors are identical is called an *ultrapower*: an ultrapower of $\mathcal{M}$, relative to the ultrafilter U, will be denoted $\Pi_U\mathcal{M}$.

The key result for ultraproducts is *Łoś's theorem*:[18] given an ultraproduct $\Pi_U\mathcal{M}_i$ and an n-ary Σ-formula ϕ, ϕ is satisfied by the n-tuple $\langle \mathbf{a}_1, \ldots, \mathbf{a}_n\rangle$ in $\Pi_U\mathcal{M}_i$ if and only if the set of indices i such that ϕ is satisfied by $\langle a_1^i, \ldots, a_n^i\rangle$ in $\mathcal{M}_i$, is in U. Thus, roughly speaking, an ultraproduct satisfies a formula if and only if "almost all" the models in the indexed family satisfy that formula. As a corollary, given some ultrapower $\Pi_U\mathcal{M}$, ϕ is satisfied by the n-tuple $\langle a_1, \ldots, a_n\rangle$ (where every $a_p \in |\mathcal{M}|$) if and only if ϕ is satisfied by $\langle \mathbf{a}_1, \ldots, \mathbf{a}_n\rangle$ in $\Pi_U\mathcal{M}$, where for every $i \in I$ and $1 \leq p \leq n$, $a_p^i = a_p$.

With this in hand, we can turn to the proof itself.[19]

Proof. The left-to-right direction of the proof is reasonably straightforward: it is just a matter of verifying that for any translation $\tau : T_1 \to T_2$, τ^* preserves isomorphisms, ultraproducts and domains.

Now consider the right-to-left direction. For any model $\mathcal{M}$ of T_2, let $\mathcal{M}^+$ be the definitional expansion of $\mathcal{M}$ to $\Sigma_1 \cup \Sigma_2$ defined by σ: that is, the model such that (i) $|\mathcal{M}^+| = |\mathcal{M}|$; (ii) for any $R \in \Sigma_1$ and any $a_1, \ldots, a_n \in |\mathcal{M}|$, $a_1, \ldots, a_n \in R^{\mathcal{M}^+}$ iff $a_1, \ldots, a_n \in R^{\sigma(\mathcal{M})}$ (exploiting the fact that σ preserves domains); and (iii) for any $S \in \Sigma_2$ and any $b_1, \ldots, b_m \in |\mathcal{M}|$, $b_1, \ldots, b_m \in S^{\mathcal{M}^+}$ iff $b_1, \ldots, b_m \in S^{\mathcal{M}}$. Now define

$$K := \{\mathcal{M}^+ : \mathcal{M} \in \mathrm{Mod}(T_2)\} \tag{7}$$

First, we want to show that K is axiomatisable. A necessary and sufficient condition for this is that K is closed under isomorphism and taking ultraproducts, and has the property that for any structure $\mathcal{A}$, if some ultrapower of $\mathcal{A}$ lies in K then $\mathcal{A}$ itself lies in K.[20] The first two properties follow immediately from σ's preservation of isomorphism and ultraproducts, so it remains only to prove the third.

[18] See, for example, (Hodges, 1997, Theorem 8.5.3).

[19] The proof below is a simplified version of the proof (of a more general result) given by van Benthem and Pearce (1984).

[20] (Hodges, 1997, Corollary 8.5.13)

To this end, suppose that $\Pi_U \mathcal{A} \in K$. Let us denote the reduct of $\mathcal{A}$ to Σ_2 by $\mathcal{M}$;[21] we want to show that $\mathcal{A} = \mathcal{M}^+$. Taking reducts commutes with taking ultraproducts,[22] so the reduct of $\Pi_U \mathcal{A}$ is $\Pi_U \mathcal{M}$. Since $\Pi_U \mathcal{A} \in K$, $\Pi_U \mathcal{M} \in \mathrm{Mod}(T_2)$. By Łoś's theorem, $\mathcal{M} \in \mathrm{Mod}(T_2)$. So $\mathcal{A}$ is an expansion of a T_2-model. It remains to show that for any $R \in \Sigma_1$ and any $a_1, \ldots, a_n \in |\mathcal{A}|$, $\langle a_1, \ldots, a_n \rangle \in R^{\mathcal{A}}$ iff $\langle a_1, \ldots, a_n \rangle \in R^{\sigma(\mathcal{M})}$. This goes as follows:

$$
\begin{array}{lll}
\langle a_1, \ldots, a_n \rangle \in R^{\mathcal{A}} & \text{iff} \quad \langle \mathbf{a}_1, \ldots, \mathbf{a}_n \rangle \in R^{\Pi_U \mathcal{A}} & \text{(Łoś's theorem)} \\
& \text{iff} \quad \langle \mathbf{a}_1, \ldots, \mathbf{a}_n \rangle \in R^{\sigma(\Pi_U \mathcal{M})} & \text{(since } \Pi_U \mathcal{A} \in K\text{)} \\
& \text{iff} \quad \langle \mathbf{a}_1, \ldots, \mathbf{a}_n \rangle \in R^{\Pi_U \sigma(\mathcal{M})} & \text{(}\sigma \text{ preserves ultraproducts)} \\
& \text{iff} \quad \langle a_1, \ldots, a_n \rangle \in R^{\sigma(\mathcal{M})} & \text{(Łoś's theorem again)}
\end{array}
$$

Thus, there is some theory T^+ such that $K = \mathrm{mod}(T^+)$. But now observe that for any two models $\mathcal{A}, \mathcal{B}$ of T^+ (i.e. any $\mathcal{A}, \mathcal{B} \in K$), if the reducts of $\mathcal{A}$ and $\mathcal{B}$ to Σ_2 are identical, then $\mathcal{A} = \mathcal{B}$. Thus, T^+ *implicitly defines* every $R \in \Sigma_1$ in terms of Σ_2. But then by Beth's theorem, T^+ explicitly defines R, and so entails some sentence of the form $\forall x_1 \ldots \forall x_n (R x_1 \ldots x_n \leftrightarrow \tau_R(x_1, \ldots, x_n))$. So, we define an interpretation $\tau : \Sigma_1 \to \Sigma_2$ by setting $\tau(R) = \tau_R$ for every $R \in \Sigma_1$, and extending by requiring τ to commute with the logical constants. It is straightforward to confirm that $\sigma = \tau^*$, and that τ is a translation from T_1 to T_2. □

Now consider the second question raised at the start of this section: whether the conditions on Theorem 4 can be thought of as imposing a requirement of similarity-preservation. One way to show this would be to argue that an ultraproduct can be regarded as a kind of "limit" of a family of models. In support of this, we could observe that in the context of propositional logic (which, recall, is simply first-order logic where all predicates are nullary) the ultraproduct of a sequence of models is the limit of that sequence in the Stone topology.[23] More generally, Łoś's theorem provides a sense in which an indexed family of models "converges" on the ultraproduct. However, the sense of convergence here must remain rather loose. The standard axioms governing convergence require that if all members of a converging family are identical to some object, then the limit of the family is that object;[24] yet, in general, the ultrapower of a given model is not identical to that model, nor is it even isomorphic to it.[25] Nevertheless,

[21] That is, $\mathcal{M}$ is the Σ_2-model such that $|\mathcal{M}| = |\mathcal{A}|$, and for any $R \in \Sigma_2$ and any $b_1, \ldots, b_m \in |\mathcal{M}|$, $b_1, \ldots, b_m \in S^{\mathcal{M}}$ iff $b_1, \ldots, b_m \in S^{\mathcal{A}}$.

[22] (Hodges, 1997, Theorem 8.5.1)

[23] Halvorson and Tsementzis (2017)

[24] See e.g. Dudley (1964), Patten (2014).

[25] Although it is, by Łoś's theorem, elementarily equivalent to it. We could try working with elementary-equivalence classes of models instead, and see if the ultraproduct does provide an appropriate notion of convergence in that setting. However, then there will be the problem of how to justify demanding

I conclude that the above considerations do illuminate some aspects of the relationship between reduction (thought of as an essentially syntactic relation between theories) and supervenience (thought of as an essentially semantic relation between models of theories, or the possible worlds they represent): we have learned that constraints on a supervenience map, statable in purely semantic terms, can enforce reducibility.

References

Barrett, T. W. and Halvorson, H. (2016). Glymour and Quine on Theoretical Equivalence. *Journal of Philosophical Logic*, 45(5):467–483.

Butterfield, J. (2011a). Emergence, Reduction and Supervenience: A Varied Landscape. *Foundations of Physics*, 41(6):920–959.

Butterfield, J. (2011b). Less is Different: Emergence and Reduction Reconciled. *Foundations of Physics*, 41(6):1065–1135.

Dizadji-Bahmani, F., Frigg, R., and Hartmann, S. (2010). Who's Afraid of Nagelian Reduction? *Erkenntnis*, 73(3):393–412.

Dudley, R. M. (1964). On sequential convergence. *Transactions of the American Mathematical Society*, 112(3):483–507.

Fletcher, S. C. (2016). Similarity, Topology, and Physical Significance in Relativity Theory. *The British Journal for the Philosophy of Science*, 67(2):365–389.

Halvorson, H. (2012). What Scientific Theories Could Not Be. *Philosophy of Science*, 79(2):183–206.

Halvorson, H. and Tsementzis, D. (2017). Categories of Scientific Theories. In Landry, E., editor, *Categories for the Working Philosopher*. Oxford University Press, Oxford. References are to draft of July 29, 2015.

Hodges, W. (1997). *A Shorter Model Theory*. Cambridge University Press, Cambridge, UK.

List, C. (2018). Levels: Descriptive, Explanatory, and Ontological. *Noûs*. Forthcoming. References are to the Early View version available at doi: 10.1111/nous.12241.

that the supervenience mapping preserve isomorphism.

Nagel, E. (1979). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Hackett, Indianapolis.

Patten, D. R. (2014). *Problems in the Theory of Convergence Spaces*. PhD dissertation, Syracuse University, Syracuse, NY.

Suppes, P. (1957). *Introduction to Logic*. Van Nostrand Reinhold, New York.

Suppes, P. (1967). What is a scientific theory? In Morgenbesser, S., editor, *Philosophy of Science Today*, pages 55–67. Basic Books, New York.

van Benthem, J. and Pearce, D. (1984). A mathematical characterization of interpretation between theories. *Studia Logica*, 43(3):295–303.

QTAIM and the Interactive Conception of Chemical Bonding

Quantum physics is the foundation for chemistry, but the concept of chemical bonding is not easily reconciled with quantum mechanical models of molecular systems. The quantum theory of atoms in molecules (QTAIM), developed by Richard F.W. Bader and colleagues, seeks to define bonding using a topological analysis of the electron density distribution. The "bond paths" identified by the analysis are posited as indicators of a special pairwise physical relationship between atoms. While elements of the theory remain subject to debate, I argue that QTAIM embodies a distinctive interactive conception of bonding that is an attractive alternative to others previously discussed.

1. Introduction

The notion of the chemical bond played a key role in the development of modern chemistry and remains central to our understanding of molecular structure and chemical transformations. However, with the advent of quantum mechanics (QM) and its successful application to molecules, it has become difficult to reconcile the traditional idea of bonds with the underlying physical theory. Some philosophers of science have grappled with this problem, discussing ways to conceive of bonds or bonding in light of QM models of molecular systems. At the same time, the quantum theory of atoms in molecules (QTAIM), a program in theoretical chemistry, provides a intriguing method for linking QM to traditional chemical concepts including bonding.

This paper is organized as follows. Section two briefly reviews the challenges facing the notion of the bond in light of modern modelling techniques, and highlights the structural and energetic conceptions discussed by Hendry (2008) and Weisberg (2008) as possible responses. An overview of QTAIM is presented in section three. While various philosophical implications of this theory have been discussed by proponents and others, I focus on the contrast it provides with these two other conceptions. In section four I argue that QTAIM provides a distinctive and appealing notion of bonding—appropriately described as an *interactive* conception. Section five briefly summarizes and suggests that the interactive conception has potential to illuminate how the idea of mechanistic explanation applies to chemistry.

2. Conceptions of the Chemical Bond in the Wake of Quantum Theory

Beginning in the 19th century, several scientists developed models of molecular structure, particularly as an avenue to explain phenomena in organic chemistry. A key figure in developing the theory of chemical bonds in the early 20th century was G.N. Lewis. Lewis (1916) distinguished two types of compounds, polar and non-polar. The former came to be described in terms of so-called ionic bonds: here electrostatic forces (which act in all directions) are responsible for the combination of oppositely charged ions. The non-polar type, Lewis reasoned, required the sharing of electrons. Specifically, the sharing of a pair of electrons between two atoms creates a covalent bond. In the theory, each element has a characteristic configuration of unpaired outer shell electrons: this is the raw material for creating covalent bonds and resulting molecules.

QM offers a very different picture of electrons, atoms and molecules. For instance, for a free particle moving in space, one cannot ascertain its position at a given time, but instead must determine its wave function Ψ, which is a function from the possible positions to a (complex) number. To interpret what this means for possible measurements of the particle's position, one calculates the probability of finding the particle in a given volume of

space at a given time from the product of Ψ and its complex conjugate Ψ* Absent a measurement interaction, the particle has no defined spatial location.[1]

For an atom, the behavior of the system is described by the time-independent Schrödinger equation,[2] $\hat{H}\psi = E\psi$, where ψ is the wave function, E is the energy, and $\hat{H}$ is the Hamiltonian operator appropriate for the system. For an atom, the Hamiltonian will contain a kinetic energy term and a potential energy term that is based on the electrostatic attraction between the electrons and the nucleus (along with repulsion between electrons). By solving the equation, one finds the wave function and the energy: in fact there are many solutions corresponding to many energy states (the lowest energy state is the ground state). In the case of the hydrogen atom (where the nucleus is assumed to be stationary at the origin of the coordinate system), the calculated wave functions (called orbitals) indicate the position state of the electron: again this is in terms of complex-valued amplitudes over the possible position configurations. For multi-electron atoms an approximate description of possible electronic states is built up from successive hydrogen-like orbitals of increasing energy. In the context of multiple electrons, one can use the wave function as the basis for calculating the *electron*

[1] The domain of the wave function for an N-particle system is a configuration space with $3N$ dimensions.

[2] The assumption required here is that the potential energy of the system does not change with time.

density distribution (ρ): this gives the expected number of electrons one would find at a particular location upon measurement.[3]

Given that electrons are not localized in quantum theory in the absence of measurement, Lewis' idea that a molecule is formed by sharing particular electrons is problematic. Linus Pauling (1960) prominently sought to reconcile the two pictures. His approach was to interpret quantum theory as describing "resonance" structures, which were hybrid combinations of multiple possible classical configurations. Some critics, however, viewed this perspective as unhelpful in the search for a purely quantum foundation for chemistry. While precise solutions to the Schrödinger equation for molecules are generally intractable, techniques to estimate a molecular wave function and thus calculate molecular energies and other properties were quickly developed.[4] One approach to calculation (which retains a conceptual link to the Lewis model and resonance theory) uses what are called valence bond (VB) models. This approach starts with the wave functions associated with

[3] For multi-electron systems, the possible position configurations will reflect that electrons are fermions and their composite wave functions are anti-symmetric. This underlies the Pauli exclusion principle, whereby only two electrons may occupy the same orbital, and they must have opposite spins.

[4] One feature which figures in all of these approaches is the Born-Oppenheimer approximation: calculations start with the assumption that the nuclei are in a fixed configuration (which can be altered iteratively to find the best solution).

individual atoms (two at a time) and creates hybrid orbitals from their overlap. As a result, VB-based calculations preserve a degree of localization in the resulting orbital structure, and the idea of overlapping orbitals provides an intuitive notion of a bond. Over time an alternative to the VB approach has become dominant: this features the use of molecular orbital (MO) models.[5] Here, one constructs orbitals for all of the electrons in the molecule together (given fixed nuclear coordinates). These orbitals are not localized: they "cover" the entire molecule. The success of MO methods for calculating molecular wave functions leads to a puzzle about how the notion of a chemical bond should be viewed given the state of the science.

In assessing this question, Hendry (2008) describes two conceptions of the chemical bond. The first, the structural conception, seeks to "retain the explanatory insights afforded by classical structural formulas (Hendry 2008, 917)." To maintain these insights in the context of quantum theory, Hendry suggests a functional approach that would identify "physical realizers" of the role traditionally played by bonds. The requirements are that the realizers would be "material parts of the molecule that are responsible for spatially localized submolecular relationships between individual atomic centers (Hendry 2008, 917)."

[5] For discussion of VB and MO models, see Weisberg (2008). Another family of models utilizing Density Functional Theory (DFT) is also frequently employed. These estimate functions on the electron density distribution to extract information about molecular properties.

Weisberg (2008) offers a slightly different definition whereby the conception says "a covalent bond is a directional, submolecular relationship between individual atomic centers that is responsible for holding the atoms together (Weisberg 2008, 934)." The main challenges facing this conception are the indistinguishability and non-localized nature of electrons in a molecule. A possible solution Hendry discusses is the identification of bonds with "nonarbitrary" components of the electronic wave function and/or electron density distribution of the molecule (Hendry 2008, 918).

The alternative Hendry outlines is called the energetic conception. Here, no part of the molecule responsible for bonding is identified. Instead facts about chemical bonding are facts about "energy changes between molecular or super-molecular states (Hendry 2008, 919)." If a molecule in a bonded state has lower energy compared to its separated atoms, this represents the formation of bonds. For two atoms forming a diatomic molecule, one can plot total potential energy as a function of inter-nuclear separation and identify the minimum value associated with bonding. For a polyatomic molecule, a potential energy surface in higher dimensions can likewise be calculated from trial wave functions. The energetic conception is, as Hendry puts it, "more a theory of chemical *bonding* than a theory of *bonds* (Hendry 2008, 919, emphasis added)." Weisberg (2008) argues that the idea that bonding involves energetic stabilization is a consistent, or robust, feature of various molecular models. This favors the energetic conception. He also argues that across the models he surveys, greater delocalization of electrons correlates with an increased match between calculated values for molecular properties and empirical estimates. This puts pressure on the

structural conception, which depends on identifying the realizers of the bond role in localized regions between atomic centers.

3. Overview of QTAIM

The quantum theory of atoms in molecules developed by Bader (1990) and colleagues offers an alternative approach to thinking about bonding.[6] The approach involves analyzing the topological features of the electron density distribution (ρ) and linking these to chemical concepts. In examining ρ for a given molecule, the most obvious characteristic is its concentration near atomic centers and low concentration elsewhere. However, a detailed examination reveals more features. Since one can treat ρ as a scalar field in three-dimensional space, one can proceed to examine the associated gradient vector field (by applying the vector differential operator ∇): this shows the direction in which the density is increasing the most at a given point (and the magnitude of the increase). In this way, one identifies features such as critical points associated with extrema (minima, maxima and

[6] Concise expositions of QTAIM include Gillespie and Popelier (2001, chaps. 6-7), and Popelier (2000, 2016).

saddle points), as well as gradient paths—trajectories that follow the line of steepest "ascent" at successive points.[7]

In examining ρ for a given molecule, a set of gradient paths originating at infinity will converge on maxima associated with each nucleus.[8] According to QTAIM, the space traversed by all of these paths (called the atomic basin), along with the nuclear "attractor" itself, defines an individual atom: "*An atom, free or bound, is defined as the union of an attractor and its associated basin* (Bader 1990, 28, emphasis original)." There is also a critical point (a saddle point) between nuclei: the set of gradient paths originating at infinity and converging on these points define a boundary, called the interatomic surface.[9] The atom is bounded inside the molecule by this surface and extends to infinity in the open directions away from the rest of the molecule: in practice the boundary in these directions may be defined using a pragmatic cut-off level of electron density (e.g. 0.001 a.u.). Next, one can

[7] Further details can also be found by examining the second differential operator or Laplacian $\nabla^2(\rho)$, which is interpreted as indicating local concentration and depletion of density. This can be used to identify (imperfect) analogues of localized electron pairs (see Bader 1990, chap. 7).

[8] Technically, these maxima are not true critical points due to discontinuities, but are treated as such as a practical matter (see Bader 1990, 19).

[9] This surface is also referred to as a zero-flux surface, in that no gradient vectors cross it at any point (see Bader 1990, 28-9).

observe gradient paths that mark out lines of concentrated density linking two atomic centers to these same inter-nuclear critical points. These are called "bond critical points" and the paths that run from it to the paired nuclei are used to define what QTAIM calls "bond paths:" the full set of bond paths comprises what is called the molecular graph (Bader 1990, 32-3).

What is the relationship between QTAIM's bond path and other notions of the chemical bond? Gillespie and Popelier (2001) caution that "a bond path is not identical to a bond in the sense used by Lewis (Gillespie and Popelier 2001, 152)." A molecular graph will not be identical to a Lewis structure, for instance, because "double and triple bonds are represented by only one bond path (Gillespie and Popelier 2001, 152)."[10] Still, they assert that "the existence of a bond path between two atoms tells us that these atoms are bonded together (Gillespie and Popelier 2001, 153)." Given this claim, one might ask if QTAIM is a way to "hold on to the structural conception of the bond understood functionally (Weisberg, Needham, and Hendry 2016, section 4.3)."

On this point, it is important to note that the QTAIM definition of a bond path also includes a stipulation that draws on the *energetic* conception of bonding. The identification of a bond path depends not only the presence of the signature pattern of electron density, but also requires that "the forces on the nuclei are balanced and the system possesses a minimum energy equilibrium internuclear separation (Bader 1990, 33)."Otherwise the feature is

[10] Part of the QTAIM approach is to look closely at the characteristics of the BCP's and neighboring topology to show how they correspond with various types of bonds.

referred to as an "atomic interaction line" (see Bader 1990, 32). But there is no reason a structural understanding of bonding cannot include this energetic component, and the QTAIM approach does at first appear to include important elements of the structural conception (as defined in Weisberg 2008): bond paths map a directional, submolecular relationship between atomic centers. With regard to Weisberg's last criterion, Bader at times seems to endorse the notion that this feature "is responsible for holding the atoms together" (Weisberg 2008, 934). He says "nuclei…are linked by a line through space along which electronic charge density, the glue of chemistry, is maximally accumulated (Bader 1990, 33)."[11] But is this line of concentrated electronic charge density literally the "glue"? While the issues here are subtle, the answer is no.

Bader says that the appearance of an atomic interaction line (AIL) between a pair of nuclei is a "necessary condition" for bonding, but its presence is a sufficient condition only when "the system possesses a minimum energy equilibrium internuclear separation" (Bader 1990, 33). It is then that the AIL is designated a bond path. In order to better understand this definition, the role of the accumulation of charge between two nuclei *in the process of*

[11] The notion that a region of electron density provides the "glue" or "cement" holding atoms together in a molecule is widespread in chemical texts, presumably for its heuristic value in some contexts (see, e.g., Loudon 1995). This provides the backdrop for Bader's comment.

achieving bonding and the meaning of the presence of a bond path *in equilibrium* must be distinguished.[12]

If one pictures separated atoms being brought closer, then bonding is "the situation obtained when the initially attractive Hellmann-Feynman forces acting on the approaching nuclei, and resulting from the accumulation of electron density associated with the formation of the atomic interaction line, vanish (Bader 1998, 7314)." Bader is referring here to the role of electrostatic forces, the reliance on which is justified by reference to the Hellman-Feynman theorem. This theorem implies that, given a wave function and associated ρ, all the forces on a nucleus in a molecule can be calculated based on classical electrostatics (see Gillespie and Popelier 2001, 134-36). As shown by Berlin (1951), this result can also be used to identify so-called binding and anti-binding regions of electron density in molecules. In the case of a diatomic system, charge density in the binding region between nuclei is seen as drawing the nuclei together, while density in anti-binding regions on the far side of the nuclei works to draw them apart (along with nuclear repulsion).[13]

Outside the equilibrium inter-nuclear separation, we can ascribe to the binding region the responsibility for a (net) attractive force. *At* equilibrium distances, all forces on the nuclei

[12] See discussion in Popelier (2000), 60-1.

[13] This electrostatic picture of how bonding is achieved has been challenged by a competing theory of the bonding process (discussed recently by Needham 2014). See Bader (2011) for a response.

are balanced. Now, one can still divide the electronic density distribution into binding and anti-binding regions (it should be noted here that the binding region encompasses much more than the line of density marked out by the bond path). But if there is no net force at work, it would be oversimplifying to say the binding region is "holding the atoms together." This would only tell part of the story, since one could say the anti-binding region also holds the atoms in place by keeping the inter-nuclear distance from compressing beyond the equilibrium separation. This point is perhaps clearer in comparison to Hendry's formulation of bonds as "material parts of the molecule that are responsible for spatially localized sub-molecular relationships between individual atomic centers (Hendry 2008, 917)." Binding and anti-binding regions both clearly play a role in defining the equilibrium inter-nuclear distance. And since these regions together encompass the entire molecule, there is no basis for concluding that the bond paths of QTAIM, despite highlighting a concentrated area of ρ, pick out a region that plays the functional role envisioned by the structural conception. Instead, the presence of the bond path at equilibrium appears to represent "forces [which] act on the nuclei for *any* displacement from their final equilibrium position (Bader 1998, 7314-315, emphasis added)."[14] As such, the bond path is a "universal *indicator* of bonding (Bader 1998, 7315, emphasis added)." According to QTAIM, then, the bond path is a *sign* that the bonding relationship exists, but it does not represent a region responsible for holding the atoms in place.

[14] See also Popelier (2000), 55-56.

Consistent with this conclusion is another departure QTAIM takes from a traditional view of bonds. Bonds are often pictured as localized *between* atoms in a molecule. However, bond paths (with the exception of the bond critical point) fall *inside* atomic boundaries. Atoms lie adjacent to one another along the interatomic surface. QTAIM offers a different picture that "requires the replacement of the model of structure that imparts an existence to a *bond* separate from the atoms it links – the ball and stick model or its orbital equivalents of atomic and overlap contributions – with the concept of *bonding* between atoms (Bader 1998, 7322, emphasis original)."

4. Interpreting QTAIM's Conception of Atoms and Bonding

Bader's distinction between "bonds" and "bonding" is in keeping with a difference Hendry identifies between the structural and energetic conceptions.[15] Given this, and given its reliance on energy equilibrium in its definition of bonding, one might ask how closely the QTAIM picture should be identified with the energetic conception. The key difference is that while the energetic conception emphasizes achieving a stabilizing minimum molecular energy, QTAIM goes further to provide an indicator that particular atoms are indeed bonded.

[15] Bader says that "a bond path is not to be understood as representing a 'bond,'" rather "the presence of a bond path linking a pair of nuclei implies that the corresponding atoms are bonded to one another (Bader 1990, 35)."

Only some pairs of atoms in a polyatomic molecule at an energy minimum are bonded, and bond paths pick these out. As a first pass, one might view the QTAIM conception as a hybrid of the conceptions considered above: while it relies on energetic considerations, its bond paths define directional relationships between atomic centers in keeping with the structural conception.

But seeing the QTAIM idea as a hybrid arguably misses what makes it distinctive. The goal of QTAIM is to provide not only a conception of bonding, but also of the bonded *atoms*: it seeks to show that atoms in molecules should be seen as bona fide physical systems in their own right. Bader notes that "quantum mechanics has been shown to account for the properties of isolated atoms and for the total properties of a molecular system" but there is a "lack of a quantum definition of an atom in a molecule (Bader 1990, 131)." The approximated solutions to the molecular wave function feature delocalized electron orbitals around a configuration of stationary nuclei. As discussed, QTAIM uses a topological examination of ρ to define atoms. In doing so, QTAIM also provides a way to calculate various atomic properties. To calculate atomic charge, for example, one integrates ρ over the topologically defined volume of an atom and then subtracts it from the associated nuclear charge: Bader argues that the consistency of these calculated values across molecules that incorporate the same atom demonstrates the success of the approach.[16] QTAIM extends this approach to other properties, including atomic energies, although this involves more complex

[16] Bader uses Li in LiF, LiO, and LiH as an example (Bader 1990, 135).

derivations.[17] Bader's ultimate claim is that QTAIM provides a full account of atoms in molecules as quantum physical subsystems: the topologically defined atom is also a quantum atom.[18] This claim continues to be the subject of debate in the theoretical chemistry literature, and no definitive judgments on its technical merits can be made here. Rather, with this sense of the goals of the program, we can return to the question of what QTAIM's approach implies for the conceptions of atoms and bonding.

Instead of describing a molecule as a system featuring interactions between electrons and nuclei, QTAIM posits atoms as interacting systems within the molecule. An atom in a molecule is an "open quantum subsystem, free to exchange charge and momentum with its environment (Bader 1990, 169)." Of course the relationship of interest is between two bonded atoms along an interatomic surface: "it is through the exchange of electrons and the fluxes in properties across the surface described by the physics of a proper open system that atoms adjust to the presence of their bonded neighbors (Bader 1998, 7322)." Bonding is a special physical relationship between pairs of atoms in a molecule where displacement

[17] Popelier (2016) gives a concise account of QTAIM's derivation of atomic energies. In response to some criticisms of the approach he concedes that it is not ruled out that some molecular fragments, which are not QTAIM's topological atoms, may also have a well-defined energy (see Popelier 2016, 37).

[18] Bader's arguments that quantum mechanical principles apply to QTAIM's atoms are given in Bader (1990), chaps. 5-6, 8.

(within limits) leads to particular restorative responses within the molecular framework.[19] In equilibrium, the nature of this relationship can be examined via the topological properties of density at the points where the atoms meet and where charge or other properties would be exchanged – the bond critical points on the interatomic surfaces. But even though bond paths are defined at equilibrium, the distinctive feature of the conception is that it embodies the idea of a particular pairwise interaction between atoms (not just between the electrons and nuclei). Rather than a combination of the energetic and structural conceptions discussed above, the QTAIM conception is better labeled an *interactive* conception of bonding.

It should be noted that QTAIM's claim of a close extensional match between its definition of bonding and traditional chemical definitions has been challenged. Bader had put the claim this way: "the network of bond paths…is found to coincide with the network generated by linking together those pairs of atoms that are assumed to be bonded to one another on the basis of chemical considerations (Bader 1990, 33)." However, Weisberg, Needham, and Hendry note that Bader's approach appears to be "too permissive (Weisberg, Needham, and Hendry 2016, section 4.3):" citing Cerpa, Krapp, Vela, and Merino (2008), they say the problem is that bond paths appear between non-bonded atoms (one example given was that of an Argon atom trapped within a C_{60} molecule which features a bond path

[19] For a challenge to QTAIM's association of bond paths/BCP's with stabilizing interactions see Poater, Solà, and Bickelhaupt (2006). The present discussion is limited to arguing that this notion is conceptually central to the theory.

connecting it with all sixty carbon atoms).[20] In recent papers, Bader clearly acknowledges that bond paths are present in contexts not traditionally associated with chemical bonds. However, he attempts to turn this into a virtue, saying QTAIM offers a more theoretically precise approach to bonding that extends beyond traditional notions but also offers analytic tools to more precisely characterize different cases (see Bader 2011, 20). On balance, while the criticism has some merit given claims made in Bader's earlier work, it must be considered in the larger context of debates about chemical bonding: neither Lewis's theory nor any successor account is free of challenges.[21]

5. Summary and Implications

The interactive conception embodied in QTAIM has advantages over the others discussed: it offers more detail about how atoms relate to one another inside a molecule compared to the energetic conception, and unlike traditional structural approaches it relies

[20] See Bader (2009) for a response. For another critique see Foroutan-Nejad, Shahbazian, and Marek (2014), who emphasize that bond paths may disappear/re-appear due solely to nuclear vibrations in some cases. They also note that atoms in molecules that are not linked by bond paths may be seen as interacting.

[21] This is true even of a minimalist energetic account. For example, Berson (2008) argues there are cases where covalent bonding leads to energetic destabilization.

only on information drawn from quantum mechanically derived calculations. It also invites one to consider that bonding is not best understood in static terms. While analyses of physical systems often center on idealized equilibrium conditions, molecules should perhaps be understood as constituted from patterns of repeated characteristic interactions between atoms in an environment of ongoing change.

This idea suggests that the interactive conception has implications for the philosophy of explanation as it pertains to chemistry. In particular, despite the successes of quantum models, approaches to mechanistic explanation typically used elsewhere in the context of complex (often biological) systems would still be applicable to molecules. First, I note that in the constitutive dimension of mechanistic explanation, systems are typically conceived of as composite entities whose properties and behaviors are due to both the properties of and the organized pattern of interaction among its constituent parts.[22] This comports well with the interactive conception of bonding. Further, the interactive view may help to elucidate the notion of mechanism in the context of chemical transformations. Goodwin (2012) describes chemists as using both a thick and thin conception of mechanism. In the thick sense of a reaction mechanism, for instance, a reaction is conceived of as a continuous evolution from reactants to products and might be represented as a path along a potential energy surface. The thin conception, on the other hand, breaks the reaction into a sequence of steps, and may feature a description of links between discrete classical structures. If atoms can be given an

[22] An example is Wimsatt's discussion of reductive explanation (see Wimsatt, 2007, 275).

energetic analysis while also being treated as interacting entities both inside and outside of molecules, this offers a potential path toward bridging these two conceptions of mechanism.

References

Bader, Richard F.W. 1990. *Atoms in Molecules: A Quantum Theory.* Oxford: Clarendon Press.

———. 1998. "A Bond Path: A Universal Indicator of Bonded Interactions." *The Journal of Physical Chemistry A* 102: 7314-23.

———. 2009. "Bond Paths are not Chemical Bonds." *The Journal of Physical Chemistry A* 113: 10391-96.

———. 2011. "On the Non-existence of Parallel Universes in Chemistry." *Foundations of Chemistry* 13: 11-37.

Berlin, Theodore. 1951. "Binding Regions in Diatomic Molecules." *Journal of Chemical Physics* 19: 208-13.

Berson, Jerome A. 2008. "Molecules with Very Weak Bonds: The Edge of Covalency." *Philosophy of Science* 75 (Proceedings): 947-57.

Cerpa, Erick, Andreas Krapp, Alberto Vela, and Gabriel Merino. 2008. "The Implications of Symmetry of the External Potential on Bond Paths." *Chemistry - A European Journal* 14: 10232-34.

Foroutan-Nejad, Cina, Shant Shahbazian, and Radek Marek. 2014. "Toward a Consistent Interpretation of the QTAIM: Tortuous Link Between Chemical Bonds, Interactions, and Bond/Line Paths." *Chemistry - A European Journal* 20: 10140-52.

Gillespie, Ronald J., and Paul L.A. Popelier. 2001. *Chemical Bonding and Molecular Geometry.* New York: Oxford University Press.

Goodwin, William. 2012. “Mechanisms and Chemical Reaction.” In *Philosophy of Chemistry*, edited by Andrea I. Woody, Robin Findlay Hendry, and Paul Needham, 309-327. Amsterdam: Elsevier.

Hendry, Robin Findlay. 2008. "Two Conceptions of the Chemical Bond." *Philosophy of Science* 75 (Proceedings): 909-20.

Lewis, Gilbert N. 1916. "The Atom and the Molecule." *Journal of the American Chemical Society* 38: 762-85.

Loudon, G. Marc. 1995. *Organic Chemistry.* 3rd. San Francisco: Benjamin-Cumming.

Needham, Paul. 2014. "The Source of Chemical Bonding." *Studies in History and Philosophy of Science* 45: 1-13.

Pauling, Linus. 1960. *The Nature of the Chemical Bond.* 3rd. Ithaca: Cornell University Press.

Poater, Jordi, Miquel Solà, and F. Matthias Bickelhaupt. 2006. "A Model of the Chemical Bond Must be Rooted in Quantum Mechanics, Provide Insight, and Possess Predictive Power." *Chemistry - A European Journal* 12: 2902-05.

Popelier, Paul. 2000. *Atoms in Molecules: An Introduction.* Harlow: Prentice Hall.

———. 2016. "On Quantum Chemical Topology." In *The Chemical Bond II*, edited by D.M.P. Mingos, 23-52. Cham: Springer International Publishing.

Weisberg, Michael. 2008. "Challenges to the Structural Conception of Chemical Bonding." *Philosophy of Science* 75 (Proceedings): 932-46.

Weisberg, Michael, Paul Needham, and Robin Hendry. 2016. "Philosophy of Chemistry." In *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2016/entries/chemistry/.

Wimsatt, William C. 2007. *Re-Engineering Philosophy for Limited Beings*. Cambridge Massachusetts: Harvard University Press.

Garson, J. (forthcoming). There are no ahistorical theories of function. *Philosophy of Science*

Title: There Are No Ahistorical Theories of Function

Author: Justin Garson

Abstract: Theories of function are conventionally divided up into historical and ahistorical ones. Proponents of ahistorical theories often cite the *ahistoricity* of their accounts as a major virtue. Here, I argue that none of the mainstream "ahistorical" accounts are actually ahistorical. All of them refer, implicitly or explicitly, to history. In Boorse's goal-contribution account, history is latent in the idea of statistical-typicality. In the propensity theory, history is implicit in the idea of a species' natural habitat. In the causal role theory, history is required for making sense of dysfunction. I elaborate some consequences for the functions debate.

Acknowledgments: I'm grateful to audience members at PSA 2018, where I presented this material. I also thank Daniel Dennett and Paul Griffiths for useful feedback.

1. Introduction

Theories of function are conventionally divided up into two main categories, historical and ahistorical (or backwards-looking and forwards-looking). The selected effects theory (Neander 1983, 1991; Millikan 1984) is an example of a *historical* theory, but there are other historical theories, including some versions of the organizational theory (McLaughlin 2001). *Ahistorical* theories include Boorse's goal-contribution account (1976; 1977; 2002), the propensity theory (Bigelow and Pargetter 1987), and the causal role theory (Cummins 1975; Craver 2001; Hardcastle 2002). In the 1970s and 1980s, it was common to see these two sorts of theories as competing with each other, though more recently, philosophers of biology have generally adopted a pluralistic stance, and see them as capturing different aspects of ordinary biological usage (Garson 2018). Still, the validity of the basic distinction has never been seriously challenged.

Many proponents of ahistorical theories have argued that we should accept their theories precisely *on account of* their being ahistorical. In other words, their alleged ahistoricity is often touted as a significant selling point of their theories, and a strong reason to prefer them over historical ones. There are two arguments along these lines. The first argument appeals to bald intuition, and says it's just obvious that functions don't always need history. One fanciful variant of this argument appeals to science fiction cases, like swamp creatures, instant lions, and randomly-generated worlds (e.g, Boorse 1976, 74; Bigelow and Pargetter 1987, 188). But one doesn't have to go as far as science fiction to find plausible cases of ahistorical functions in biology. Many philosophers have a strong intuition that, the very first time a new biological trait emerges and begins to benefit the organism, it has a *function* even if it was never selected for (e.g., Boorse 2002, 66; Bigelow and Pargetter 1987, 195; Walsh and Ariew 1996, 498). The second argument, which is closely related, appeals to ordinary biological usage instead of intuition. It says that historical theories run against the way biologists ordinarily think and talk about functions. At least sometimes, when biologists attribute functions to traits, they neither *cite* nor *refer to* nor *think about* history or evolution (e.g., Godfrey-Smith 1993, 200; Amundson and Lauder 1994, 451; Walsh 1996, 558; Boorse 2002, 73). Hence, ahistorical theories capture important strands of real biology.

In light of the above, my thesis might come as a bit of a shock. I claim that *there are no ahistorical theories of function* – or, put more precisely, the mainstream versions of the allegedly ahistorical theories on the market aren't actually ahistorical. If we poke and prod at those theories a bit, a historical element falls out, like contraband stashed away in a suitcase. In Boorse's version of the goal-contribution account, history is explicitly embedded in his notion of a *statistically-typical* contribution to fitness. In the propensity account, history is embedded, a little less explicitly, in the idea of a species' *natural habitat*. Finally, the only way the causal-role theorist can hope to make sense of dysfunction is to appeal to history.

Before I move on, there is one big qualification I must get out of the way. One could *invent* a purely ahistorical theory of function. One could assert, for example, that *all* of a trait's effects are its functions. In fact, the biologists Bock and von Wahlert (1965, 274)

proposed a theory of function very much along these lines. This theory (pan-functionalism?) would be ahistorical, to be sure, since even if the world were created two seconds ago in pretty much its present form, things would still have effects, and so they'd still have functions. In fact, sometimes scientists actually *do* use the word "function" synonymously with "effect." They say things like, "climate change is a *function* of deforestation," or "poor academic performance is a *function* of malnutrition." But this isn't the ordinary biological use, which the theories I cite above are trying to capture. I'll come back to this point in the conclusion.

So, I need to amend my thesis slightly. Instead of saying that there are no ahistorical theories of function, I want to say that any theory of function that satisfies two very minimal, very traditional, and largely uncontroversial, adequacy conditions, is *also* a historical theory. First, the theory should capture some distinction between functions and accidents (the function of the nose is to help us breathe but not hold up glasses). Second, the theory should capture the possibility of malfunctioning or dysfunction. If my heart seizes up due to cardiac arrest, it's failing to perform its function or it's dysfunctional. All of the theorists I engage with in this paper purport to satisfy these two adequacy criteria, or something in their vicinity, so I'm not begging any questions by insisting on these conditions.

Here's the plan for the rest of the paper. The next three sections will examine Boorse's goal-contribution theory, the propensity theory, and the causal-role theory, in turn. In the conclusion, I'll draw out the big consequences for thinking about functions.

2. Boorse's Goal-Contribution Account

Boorse's view (1976; 1977; 2002), at the most general level, is a goal-contribution account. It holds that a trait's function is just its contribution to a goal. Here, I'll focus on the subclass of functions he calls *physiological* functions. For Boorse, the *physiological* function of a trait is its species-typical contribution to the survival and reproductive prospects of an organism (1977, 555; 2002, 72). (To be more precise, Boorse carves up species into subgroups based on age and sex; the function of a trait is its typical contribution to fitness within the members of that subgroup.) Though he doesn't define a corresponding notion of *dysfunction*, he defines a closely related notion of *disease*: a disease is simply a state that "reduces one or more functional abilities below typical efficiency (1977, 555)."

Neander (1991, 182) raised a now-famous objection against Boorse; she pointed out that Boorse's view, as it stands, can't make sense of pandemic dysfunction: "dysfunction can become widespread within a population…A statistical definition of biological norms implies that when a trait standardly fails to perform its function, its function ceases to be its function; so that if enough of us are stricken with disease (roughly, are dysfunctional) we cease to be diseased, which is nonsense." Pandemic dysfunctions, moreover, don't just occupy the realm of science fiction, as in P. D. James' *The Children of Men*. UV radiation poisoning in anurans is a good example of pandemic dysfunction. Sadly, climate change might create many more pandemic dysfunctions very soon. A good theory

of function shouldn't close off the possibility that all, or most, tokens of a certain trait in a certain species are dysfunctional (or as Boorse prefers, "diseased").

Intriguingly, Boorse doesn't deny the possibility of pandemic disease. Instead, he says that in order to make sense of pandemic disease, one has to appreciate function's *historical depth*. Specifically, when we consider what's "statistically typical" for a trait, we cannot just look at what is typical right now. We have to examine the trait's behavior over a slice of time that includes the present moment and reaches far back into the past: "*Obviously*, some of the species' history must be included in what is species-typical (2002, 99; my emphasis)." He tells us that this time-slice should be longer than "a lifetime or two," and might include "millennia."

This is an extraordinary admission, given that much of Boorse's core argument for his view was propped up on the claim that both biology and intuition need purely ahistorical functions, uncluttered by history. His admission implies that his two key arguments for the view don't work. First, by his own lights, it's not the case that biologists don't refer to history; implicitly, when they talk about what's statistically-typical, they *are* talking about history. Second, regardless of whether or not intuition supports ahistorical functions, Boorse's theory doesn't. It's just not true, on Boorse's account, that if lions popped into being from an unparalleled saltation, their distinctive parts and processes would have functions. They wouldn't, since they don't have the right history (or to be more precise, they have no history at all).

3. The Propensity Theory

Bigelow and Pargetter (1987) also developed an influential "ahistorical" theory of function, the propensity theory. They reject the selected effects theory (and etiological accounts more generally) because the selected effects theory gets the *modality* of functions wrong. In other words, the statement, "functions are selected effects," if true, is contingently true; it might be true on the actual world, but there are possible worlds at which it's false. To illustrate the point, they ask us to consider a world that is pretty much the same as ours except that it randomly popped into being five minutes ago. On that world, they claim, there would still be functions, just no selected effects (188): "we have the intuition that the concept of biological function…[is] not thus contingent upon the acceptance of the theory of evolution by natural selection." This consideration prompts the need for an ahistorical theory.

For Bigelow and Pargetter, functions are propensities, or probabilistic dispositions. We might quibble over what exactly dispositions are, but any good definition will cite three parts: structure, environment, and behavior. Consider the solubility of salt. There is a *structure*, namely, the polar molecular structure composed of sodium and chloride; there is an *environment*, namely, water; there is a *behavior*, namely, dissolving. When we say that salt is disposed to dissolve in water, we're saying that, if you were to take something with this structure, and put it in this environment, it would perform this behavior, all things equal.

Functions, too, are dispositions. Consider "the function of the heart is to circulate blood." For this statement to be true, there must be a structure (the heart, embedded the right way in the circulatory system), an environment (which they call the creature's *natural habitat*), and a behavior (conferring a fitness boost on the organism). If one were to put the structure in its natural habitat, it would increase the fitness of the organism (relative, I suppose, to creatures without hearts). The crucial distinction between their view and Boorse's is that in their view, a trait's function doesn't depend on actual frequencies of performance. A trait needn't have an actual track record of boosting fitness to have a function; a mere propensity will do.

This raises the thorny question of what a creature's *natural habitat* is. For they're clear that a creature's natural habitat isn't just any environment the creature happens to find itself in. Curiously, they refuse to define this crucial notion; instead, they brush it off as vague, but unproblematically so: "there may be room for disagreement about what counts as a creature's 'natural habitat;' but this sort of variable parameter is a common feature of many useful scientific concepts" (192). But one could at least form the suspicion that if one analyzed this unproblematically vague notion, one would find some reference to history tucked away inside of it.

This suspicion is confirmed in the very next paragraph of their paper. There, they tell us that, if a creature's environment were to change very suddenly, then "natural habitat" will still refer to the *old* environment, and not the *new* one (ibid). There's a time lag built into the very idea of a natural habitat. So, for example, if climate change melts enough Arctic ice, then, at least for a time, the polar bear's natural habitat (and by extension, the natural habitat of the trait itself, namely, their thick, water-repellant fur) is the icy habitat of yore and not the contemporary, denuded one. They take that as given, and I agree.

But why would this be? What *makes it the case* that, in cases of rapid habitat change, "natural habitat," at least for a time, refers to the old environment and not the new one? What makes it true, I suspect, is that the idea of a natural habitat is an intrinsically historical notion. It's something like *the environment within which the species recently survived and thrived*. And if that's not what a natural habitat is, I would like to know what it is *such that*, if a creature's actual habitat shifts suddenly, the natural habitat, for a little while, is still the old one. Just because a concept is vague around the edges, that doesn't excuse one from the obligation to give some sort of analysis. Perhaps one could revise the theory and drop all reference to "natural habitat," as suggested by Griffiths (2009, 27), but that remains to be worked out in a rigorous way. Moreover, it's not clear whether such a theory, when rigorously developed, would hang together with the two adequacy criteria.

Hence, I conclude that, contrary to rumor, the propensity theory is not an ahistorical theory, or not demonstrably so. But if that's right, proponents of the propensity theory lose one of the main virtues of the view, which is to get the modality of functions right. To be fair, there's still a sense in which their view *is* ahistorical. What they can do, that the selected effects theorist can't, is to attribute functions to novel traits – so long as that

novel trait belongs to the members of a species that has been around long enough to have a natural habitat. Suppose a gene mutation confers a benefit on an organism, say, pesticide resistance in a flour beetle. I suppose they can say that, at the very moment at which it first confers that benefit, the gene mutation has a function, namely, to make the beetle withstand a certain pesticide. This result, they claim, is "intuitively comfortable" (195). But they can say that only because flour beetles themselves have a history, and so we can talk meaningfully about their natural habitats. Moreover, I think they'll still have a rough time explaining dysfunction (Neander 1991, 183), for reasons I'll point to in the next section. Finally, I think there are good theory-neutral reasons for saying that beneficial traits, on their very first appearance, don't have functions, but rather, whatever benefit they bring is a lucky accident. But I won't argue for that here (see Garson 2019, Chapter 2).

4. The Causal Role Theory

What about the causal role theory of function? This appears to be a purely ahistorical view. The causal role theory says, roughly, that the function of a *component* of a system consists in its contribution, in tandem with the other components, to a system-level capacity of interest (Cummins 1975; Craver 2001; Hardcastle 2002). Craver (2001) helpfully elaborates this view by specifying that the part in question must be a component of a *mechanism*. All of the basic ingredients of this theory, it seems, are ahistorical: capacities, components, organization, hierarchy, interests. Even if the world were created five minutes ago, in pretty much its present form, things would still have causal role functions.

The problem enters when we think about dysfunction. Cummins (1975, 758) insisted that functions are dispositions, or capacities: "…to attribute a function to something is, in part, to attribute a disposition to it." The function of a trait *token*, then, consists in its capacity to contribute to a system-level effect. But what if the token in question, through defect or disease, loses the capacity, and so can't contribute to the system-level effect? Then, by Cummins' analysis, it doesn't have the relevant function – so it can't be dysfunctional either.

Causal role theorists have, by and large, been silent about how to make sense of dysfunctions from this perspective. Almost everything they've had to say on that score, however, is consistent with the following theme: a trait *token* is dysfunctional when it can't do what other trait tokens generally, or typically, do to contribute to the system-level effect of interest. Consider Godfrey-Smith (1993, 200): "Although it is not always appreciated, the distinction between function and *mal*function can be made within Cummins' framework…If a token of a component of a system is not able to do whatever it is that other tokens do, that plays a distinguished role in the explanation of the capacities of the broader system, then that token component is malfunctional." Craver (2001, 72), offers the same general line: "…the ascription of a function to a malformed or broken part is derivative upon a description of how that *type* of part (X) fits into a *type* of higher-level mechanism (S). The malformed and broken part can be identified as an X by

the typical properties and activities of Xs....” This is, at root, to rely on a statistical norm for making sense of dysfunction.

This account of dysfunction, like Boorse’s, stumbles when it encounters the problem of pandemic dysfunction. For the modification suggested above implies that, if everyone’s heart seized up at once, nobody’s heart would have a function anymore, so nobody’s heart would be dysfunctional. The best way to solve this problem, and perhaps the only way, is the way Boorse took, namely, to say that the function of a trait is its typical contribution to some system effect, where what’s typical is assessed over a chunk of time that stretches back into the past, for at least “a lifetime or two,” and perhaps “millennia.” But if causal role theorists take that line, they’d have a historical theory.

Craver (2001) and Hardcastle (2002) suggest, all too fleetingly, a different way of thinking about dysfunction, one that depends not on statistics, but on our values, that is, the values and goals of people who make function attributions. Craver (2001, 72) suggests that traits are dysfunctional when they cannot do what people *want* them to do: “the mechanistic role of the broken part only appears against the fixed backdrop of shared assumptions about a type of mechanism within which parts of this type generally (or preferably) make important contributions.” The parenthetical remark alludes to a substantially new doctrine, one that demands our full concentration. It suggests that dysfunction is a mirror of human preferences and goals, of our wishing and wanting. If my heart seizes up, it’s dysfunctional, since it’s not doing *what I want it to do*.

Hardcastle (2002) makes remarks along similar lines. She first says that the function of a trait – what it’s “supposed to do,” as she puts it – depends on the goals of the scientific discipline that makes the investigation: “The teleological goal for some trait…depends upon the discipline generating the inquiry” (153). The palmomental reflex causes a chin twitch when you stroke an infant’s palm; it’s just an accident of cortical wiring with no deep evolutionary rationale. Still, she says, it has the *function* of indicating the state of brain development in infants, because that’s how biomedical researchers use it. She then says that something is malfunctional just when it cannot do what it’s supposed to do (152). The palmomental reflex is malfunctional when it can’t indicate the state of brain development. Simply put, dysfunction happens when a trait can’t do what we want.

But dysfunctions can’t be reduced to mere preferences in any straightforward way; this is a point that’s been taken in the literature for decades (e.g., Boorse 1977, 544; Wakefield 1992, 372), for reasons that scarcely need to be rehearsed. I’d prefer not to need sleep and water; I’d prefer if nobody had to go through the pain of childbirth or teething, either. But none of those things are dysfunctions. For that matter, I’d prefer if my hands were equipped with retractable adamantium claws. The fact that my hands can’t do what I want them to do doesn’t make them dysfunctional. If one really wanted to run with this value-centered line about dysfunction, one would *at least* have to add that, in order for a trait to be dysfunctional, it’s not enough that it doesn’t do what I prefer, but I must also have a *reasonable expectation* that it *should* act in the way that I prefer. But what could possibly ground a *reasonable expectation* that my hand (say) work in a certain way? Only this: that hands usually *do* work in the preferred way. But then we’re back to statistical

norms, and long historical slices of time. This value analysis of dysfunction isn't a contender to a statistical analysis; instead, the former presupposes the latter.

I've walked through three allegedly ahistorical theories of function, and shown that none of them are purely ahistorical; they're *infected* with history. The conclusion will say what we should do next.

5. Conclusion

There are no ahistorical theories of function, at least among the mainstream theories that are put forward as ahistorical. The first, Boorse's goal-contribution theory, explicitly refers to what's statistically typical for a trait, where what's typical is assessed over a long historical period of time. The second, the propensity theory, refers to the creature's natural habitat, which is implicitly historical. And the third, the causal role theory, can't hope to make sense of dysfunction (or so I argue) without appealing to a statistical norm, and thereby (following Boorse) to history. None of these theories will give functions to the parts of swamp creatures, instant lions, or anything on worlds that are similar to ours except for being randomly generated five minutes ago. The propensity theory, at least, can give functions to novel traits as soon as those traits begin benefiting their bearers, as long as the population in which the traits emerge has been around for long enough to have something like a natural habitat. But even that theory will probably encounter problems when it comes to making sense of dysfunction, though I haven't pushed that line in any detail here.

If my thesis is correct – that there are no ahistorical theories of function – three consequences immediately follow. First, we need to jettison this whole way of dividing up theories of function. The distinction between etiological and non-etiological theories serves us much better. An *etiological* theory holds that function ascriptions either are, or purport to be, causal explanations for the existence of traits. Non-etiological theories hold that function ascriptions are not, and they don't purport to be, causal explanations for traits. But the crucial point is that being etiological and being non-etiological are just *two different ways of being historical*.

Second, given that there are no ahistorical views, the two main arguments that have repeatedly been put forward for those theories – the argument from intuition and the argument from ordinary biological usage – don't actually work. If we took those arguments seriously, they'd count as evidence *against* these allegedly ahistorical theories. That doesn't mean those theories are wrong. It does mean, however, that we need to rethink, from the ground up, the motivation for accepting those theories.

A third consequence is that one popular way of thinking about function pluralism must fail. This sort of pluralist wishes to sort all biological usage under two main umbrella theories, the selected effects theory and the causal role theory. An argument for this sort of pluralism is that it mirrors the two main uses of "function" in biology, the historical sense and the ahistorical sense. If I'm right, this incarnation of the pluralist project can't work either.

True, there are some theories of function I haven't addressed here, which fall a bit outside of the mainstream. Might those come to our rescue? In particular, one might wonder how the *modal theory* of function (Nanay 2010) fares with respect to my analysis. The modal theory holds that a function of a trait *token* depends on that token's behavior on nearby possible worlds, where what's "nearby" depends on our explanatory interests. I agree that this is an ahistorical theory through and through, since what function a trait has, and whether or not it's dysfunctional, depend on what's going on at other possible worlds, rather than the actual past. But it also yields a deeply implausible construal of dysfunction. As Neander and Rosenberg (2012) point out, if the modal theory is right, then many traits that biologists don't think of as dysfunctional, like the trait of lactose-intolerance in most Pacific Islanders, would actually be dysfunctional. So, while the modal theory doesn't violate the *letter* of my second adequacy condition – namely, that it should allow for the possibility of dysfunction – it violates the *spirit* of that condition by carving up functions and dysfunctions in a wildly revisionary way.

Nanay (2012) argues that the fact that function ascriptions are relative to our explanatory interests can somehow lessen the sting of this counterintuitive consequence, but I don't see how that helps. To illustrate the problem, consider Temitope, an evolutionary geneticist who's interested in how human beings might evolve in the near future. Temitope considers a possible world to be "nearby" if, at that possible world, she has a counterpart, and her counterpart's genome differs from hers by only a single point mutation, but the rest of the world is largely the same (yielding at least 3 billion nearby worlds). She reasons that, on some of those possible worlds, some of her traits would do things that enhance her inclusive fitness. For example, we might suppose that there is a possible world at which her body's ability to dissolve arterial plaque is substantially enhanced, one at which she has tetrachromatic vision, and one at which she's resistant to malaria. She realizes, with dismay, that her body's actual ability to dissolve arterial plaque represents a dysfunction. In fact, she realizes that, *relative to her explanatory interests*, she has many more dysfunctions than she ever thought possible. So even if we agree that function ascriptions are tethered to explanatory interests, we still get deeply revisionary consequences. In my reckoning, a theory that hangs together pretty well with ordinary biological usage is better than a deeply revisionary one, all things equal (see Garson 2016, 105-7, for further discussion).

There's a twist to my story, which I alluded to in the introduction. I think there is a prominent sense of "function" in scientific circles that is ahistorical. Consider that climate change is a function of deforestation, poor academic performance is a function of malnutrition, and wildlife habitat is a function of soil. These notions are ahistorical through and through. "Function," in this context, means little more than "effect," and perhaps (as in the last of the three examples) "helpful effect." But this tepid sense of function isn't going to sustain a distinction between function and accident, nor will it give us any sense of dysfunction. This is the sort of "function" that Bock and von Wahlert (1965, 274) were getting at when they equated functions with "all physical and chemical properties arising from [the trait's] form." It's also the sort of "function" that Neander (2017) describes in her recent discussion of "minimal functions." But the proponents of

the allegedly ahistorical theories want functions to do much more than that. They are trying to capture the ordinary biological sense (or *an* ordinary biological sense) of "function," where functions differ from accidents and sometimes things are dysfunctional. Unfortunately, they can't have what they want.

References

Amundson, R., and G. V. Lauder. 1994. Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy* 9: 443-469.

Bigelow, J., and Pargetter, R. 1987. Functions. *Journal of Philosophy* 84: 181-196.

Bock, W. J., and von Wahlert, G. 1965. Adaptation and the form-function complex. *Evolution* 19: 269-299.

Boorse, C. 1976. Wright on functions. *Philosophical Review* 85: 70-86.

Boorse, C. 1977. Health as a theoretical concept. *Philosophy of Science* 44: 542- 573.

Boorse, C. 2002. A rebuttal on functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M. Perlman, 63-112. Oxford: Oxford University Press.

Craver, C. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72: 741–765.

Garson, J. 2016. *A Critical Overview of Biological Functions*. Dordrecht: Springer.

Garson, J. 2018. How to be a function pluralist. *British Journal for the Philosophy of Science* 69: 1101-1122.

Garson, J. 2019. *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press.

Godfrey-Smith, P. 1993. Functions: Consensus without unity. *Pacific Philosophical Quarterly* 74: 196-208.

Griffiths, P. 2009. In what sense does 'nothing make sense except in the light of evolution'? *Acta Biotheoretica* 57: 11-32.

Hardcastle, V.G. 2002. On the normativity of functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M Perlman, 144-156. Oxford: Oxford University Press.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Nanay, B. 2010. A modal theory of function. *Journal of Philosophy* 107: 412-431.

Nanay, B. 2012. Function attribution depends on the explanatory context: A reply to Neander and Rosenberg's reply to Nanay. *Journal of Philosophy* 109: 623-627.

Neander, K. 1983. *Abnormal Psychobiology*. Dissertation, La Trobe.

Neander, K. 1991. Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.

Neander, K. 2017. Functional analysis and the species design. *Synthese* 194: 1147-1168.

Neander, K., and Rosenberg, A. 2012. Solving the circularity problem for functions. *Journal of Philosophy* 109: 613-22.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47: 373–388.

Walsh, D.M. 1996. Fitness and function. *British Journal for the Philosophy of Science* 47: 553-574.

Walsh, D. M., and A. Ariew. 1996. A taxonomy of functions. *Canadian Journal of Philosophy* 26: 493-514.

Wright, L. 1973. Functions. *Philosophical Review* 82: 139-168.

Time-sensitivity in Science

Daria Jadreškić

Abstract

I examine the role of time-sensitivity in science by drawing on a discussion between Kevin Elliott and Daniel McKaughan (2014) and Daniel Steel (2016), on the role of non-epistemic values in theory assessment and the epistemic status of speed of inference. I argue that: 1) speed supervenes on ease of use in the cases they discuss, 2) speed is an epistemic value, and 3) Steel's account of values (2010) doesn't successfully distinguish extrinsically epistemic from non-epistemic values. Finally, I propose an account of time-sensitivity.

1. Introduction

Kevin Elliott and Daniel McKaughan (2014) argue that non-epistemic values sometimes legitimately take priority over epistemic ones in assessing scientific theories, models, and hypotheses because scientific representations are not only evaluated based on their fit with the world, but also based on the fit with the needs of their users. Their argument draws on accounts of scientific representation by Ronald Giere and Bas van Fraassen, and two examples: expedited risk assessments of the toxicity of substances (Cranor 1993, 1995) and rapid assessment methods for wetland banking (Robertson 2004, 2006). The examples attempt to show that non-epistemic values such as speed in the toxicity case and ease of use in the wetland banking case can have a more decisive role than that of being secondary considerations when epistemic values alone don't suffice to decide which representation to choose.

In a comment on their paper, Daniel Steel (2016) argues that both examples fail to show that epistemic values have been overridden by non-epistemic ones, but are rather cases in which non-epistemic values serve as secondary considerations for resolving epistemic uncertainty. According to Steel, the cases in question are not examples of accepting an epistemically inferior option because the argument rests on two problematic implicit premises: that it is epistemically better to wait for results generated by a more reliable method if one exists (E_1), and that it is bad from an epistemic perspective to select a simpler, less detailed model over one that is more complex and more detailed (E_2). In fact, in his (2010) article Steel uses Cranor's analysis to argue that non-epistemic values can influence scientific inferences without compromising epistemic ends. The problem he

identifies with Elliott's and McKaughan's account is that E_1 overlooks the epistemic costs of extended suspension of judgment and therefore "threatens to entail the absurd result that scientists should never accept any claim" (Steel 2016, 610) while E_2 violates the principle of Ockham's razor. Since there are many epistemic purposes to which hypotheses can be put, some of which can favor simplicity, there is nothing epistemically wrong with choosing a simpler option. Moreover, Steel characterizes both cases as illustrative of time-sensitivity*:*

"Both illustrate what I will call *time-sensitivity*, wherein it may be better for practical or social reasons to accept the results of a quicker-but-less-reliable method rather than wait for a slower-but-more-reliable-one. In both instances, there is a pressing interest to draw inferences in a timely manner: the protection of public health in the first and the economic interest of not unduly delaying construction projects in the second." (Steel 2016, 609)

My aim in this paper is to examine the role of time-sensitivity is science. I start by arguing against Elliott's and McKaughan's view that the two tokens, speed and ease of use, independently of one another represent the same type, namely a non-epistemic value that sometimes takes priority over epistemic ones in assessing scientific representations. Besides the problem of labeling speed and ease of use as non-epistemic, I claim that in both cases speed supervenes on simplicity and ease of use, i.e. the methods are simple and easy to use in order to be fast and enable fast (soon and many) applications. Both case studies are in fact primarily about speed, as already the titles of Elliott's and McKaughan's chapters reveal: *Expedited* Risk Assessments and *Rapid* Assessment Methods.

In the third chapter I argue that speed is an epistemic value, contrary to Elliott and McKaughan and closer to Steel, but I part from the latter in that I don't think that the epistemic/non-epistemic distinction suffices for explaining decision making in science.

I proceed by examining a way to account for time-sensitivity with the help of Steel's conceptual framework. He offers a version of epistemic values which purports to argue in favor of maintaining the epistemic/non-epistemic distinction, as well as to be useful for delineating legitimate from illegitimate influence of non-epistemic values in research, namely by distinguishing between extrinsically and intrinsically epistemic values. (Steel 2010) It seems to be consistent with Steel's account to consider time-sensitivity an extrinsic epistemic value, since he argues for a broad understanding of epistemic values: "Epistemic values can be manifested by things other than theories and hypotheses, such as methods, social practices, and community structures." (2010, 19) In this case, time-sensitivity might be a value manifested by social practices. However, I show that Steel's account of values doesn't prove to be helpful for handling the epistemic/non-epistemic controversy because it fails to distinguish between extrinsic epistemic values and non-epistemic values, especially when their influence on scientific research is legitimate, i.e. when they don't obstruct the attainment of truth.

In the fourth chapter, I claim that time-sensitivity isn't captured well in either of the contrasting notions of value distinctions. I argue that time-sensitivity is not a value of methods, but of problems to be solved in their particular contexts. We implicitly or explicitly assign a degree of time-sensitivity to problems in their specific contexts, a value judgment about when we want or expect to have results from a particular instance of

research, but it is neither a value exclusively external nor internal to science, but a requirement of efficiency which is both truth seeking and temporally constrained.

2. Speed Supervenes on Ease of Use and Simplicity

The first example presented in Elliott's and McKaughan's paper is based on Carl Cranor's analysis (1993, 1995) of different modelling approaches for assessing risks posed by toxic substances that are not pesticides or pharmaceuticals. In the United States the burden of proof is on the government to show that these products should be restricted or removed from the market and not on the manufacturers that produce them. Cranor analyzes trade-offs between different modelling approaches for assessing risks and concludes that social costs of relying on risk-assessment procedures which are rather accurate but slow are greater than of less accurate but quicker methodologies. This conclusion is based on the case of California Environmental Protection Agency (CEPA) which used an expedited risk assessment methodology in the early 1990s and was able to estimate carcinogenic potency of 200 chemicals in an 8 month period, while the traditional methodology was able to assess only 70 chemicals in 5 years, though with greater accuracy. The expedited procedure is called the linearized multistage default method (LMS) – it uses a carcinogenic potency data base, State of California data selection procedures and state-mandated default assumptions to facilitate otherwise time-consuming and science-intensive tasks in estimating dose-response relationships. Cranor calculates the difference between false positives and false negatives using different estimates, some more and some less favorable

to the expedited approach. It turns out to be a better approach in every case, in terms of minimizing social costs connected to under-regulation of likely carcinogens. Elliott's and McKaughan's conclusion is that speed is in this case prioritized over accuracy.

The second case deals with Rapid Assessment Methods (RAMs) for assessing similarity between different wetlands as part of mitigation measures when damaging or drying wetland areas. A destroyed wetland has to be compensated by preserving or restoring another wetland area, and regulatory agencies have to decide whether the destroyed and preserved wetlands are sufficiently similar so that the two could be traded. In recent years a mitigation "banking" system is developed by regulatory agencies, developers and entrepreneurs to handle mitigation. Geographer Morgan Robertson (2004, 2006) analyzes different methods to show how the banking method differs from the methods one would use if the goal was a detailed ecological characterization. Developers purchase mitigation "credits" from specialists who create "banks" of preserved or restored wetlands, in which they focus on specific features that are considered relevant for establishing the classification of 'equivalence' between wetlands. RAMs consist of algorithms that convert data about a wetland into a numerical score that estimates a wetland's functional value and is typically represented by one main score rather than a variety of different scores "in order to keep the process simple." (Elliott and McKaughan 2014, 13) This case is supposed to be illustrative of ease of use as a value that is here taking priority over predictive accuracy. Their overall conclusion is that non-epistemic values sometimes take priority over the epistemic ones.

Against this, I argue that in these two cases, we are misled to judge speed and ease of use on a par with each other, as two tokens of the same type (a non-epistemic value that trumped predictive accuracy in assessing scientific representations), when in fact we have two cases of favoring an expedited outcome, which supervenes on ease of use.[1] Speed of inference is a value that has a decisive role of taking priority over predictive accuracy, if one wants to agree that this is what happens here, while ease of use and simplicity have only a transitive role as a means to achieve faster outcomes and applications. I don't imply that speed is always dependent on ease of use or that the benefits of ease of use and simplicity reduce to speed, but I claim that this is what is going on in the two examples. For example, a theory can be simple and easy to use, but it can hardly be fast. It would be strange to claim that Euclidean geometry is faster than non-Euclidean geometry, or that Newtonian mechanics is faster than quantum mechanics. However, here we are not dealing with theories, but rather with methods and scientific practices that use simplifications, defaults, and idealizations, designed to be applied to problems in particular contexts, and these methods and practices will most likely have simplicity and ease of use contributing to speed.

Elliott and McKaughan explicitly set out to show how non-epistemic values sometimes trump the epistemic ones such as predictive accuracy, and values that have supposedly done so are speed and ease of use. Although the second example is about making wetland

[1] To some extent, ease of use of the method supervenes on its simplicity, but this relation is not of our interest at the moment.

models easy to use, rather than being highly accurate, the reason for doing this is to make them readily available and thus – faster to use. RAMs or 'rapid assessment methods' are indeed called precisely like that, but still the argument put forward is that ease of use is the value that took priority over accuracy in this case. It is certainly a feature of the method in comparison to more sophisticated ones, but Elliott and McKaughan decided to talk about non-epistemic values in general based on the sample of two values which on the closer look turn out to be cases in which one value supervenes on the other, and that is speed supervening on ease of use, and transitively also on simplicity.

We can see the connection between simplicity, ease of use, and speed in both cases. Expedited risk assessment methodology is less science- and time-intensive, RAMs are easy to use because they are simple, and therefore the results are generated faster than it would be with methods more detailed, complex, or difficult to handle. Methods do not generate results faster in order to be easy to use but are rather easy to use in order to generate results faster. It is clear that being easy to use and being fast doesn't mean the same, but easy is *here* rather to be fast, than the other way around.

3. Speed as an Epistemic Value

The status of speed of inference is disputed in the discussion. Elliott and McKaughan claim that speed is a non-epistemic value: "The cases discussed in the following sections focus on conflicts between the epistemic value of accurate prediction versus non-epistemic values such as ease of use or speed of generating results." (2014, 7) In his comment, but

also in an earlier article, Steel argues that speed is an epistemic value: "The trade-off between the speed and reliability of scientific methods, therefore, is a trade-off between two epistemic values." (2010, 27)

First of all, not everything in science that we usually attribute values to can have the value of speed. Theories and hypothesis can't be fast, but methods, applications, and more broadly, practices, can. Methods, together with theories, models, hypothesis (representations) constitute practices in science, and practices can trade off speed and accuracy depending on their applications to problems in certain contexts. Speed, together with ease of use, is therefore a feature of methods and broader, a feature of practices as applied to problems in contexts. Problems, unsurprisingly, need to be solved, so the efficiency of methods and practices becomes important and has a bearing on the balance between values internal to the scientific practice that addresses them. Steel's distinction between epistemic "building blocks" and epistemic "endpoints" is useful here. Basic science is a building block for future research so it has a slower and more cautious approach when it comes to balancing reliability and speed of inference, because an error in that context is more likely to have damaging effects by leading to more errors. In contrast, scientific results that "are used primarily for some practical purpose, such as setting allowable exposure levels to toxic chemicals or predicting climate trends (…) are more like scientific endpoints than building blocks for future knowledge" (Steel 2010, 27).

Speed of inference is an internal value of scientific research – there is always a certain speed at which methods and practices operate. We might be tempted to call it non-epistemic because motivations to prioritize speed often come from outside of science and

can operate on expense of accuracy. But when speed is understood as speed of getting at *true,* or *approximately* true results, then it has a clearly epistemic role because it moves us *temporally* closer to truth, i.e. it enables us to get in the possession of knowledge earlier and therefore advances our epistemic status. (See Steel 2016, 610) The non-epistemic part is still confined to different social and pragmatic reasons such as protection of health or economic benefits that instrumentalize speed for their reasons on expense of accuracy. However, speed is often a means to promote those without epistemic costs, as Steel argues. When it does so, the influence of those non-epistemic reasons is legitimate, when, in contrast, speed promotes them without appropriate consideration of accuracy, its prioritization, together with their influence, is illegitimate.

The source of influence is still social, pragmatic, non-epistemic, and speed itself, as a feature of a method or a practice, belongs to the internal part of science all along the way and has to be traded off against other epistemic values in any case. If non-epistemic reasons push the research in a direction that moves it away from the truth, they can distort the balance between different values, for example illegitimately prioritize speed of getting at *any r*esults over accuracy, but it can also happen that their influence on the trade-off is harmless or even beneficial, as I will explain later. Social reasons are the non-epistemic part here, not the speed that they instrumentalize.

4. Extrinsically Epistemic Equals Non-epistemic-but-legitimately-influencing

Steel's notion of epistemic values (2010) defines epistemic in terms of either intrinsically or extrinsically promoting the attainment of truth. Moreover, it allows that epistemic values are manifested by methods, social practices, and community structures. A value that Steel analyses at length as an example of an extrinsic epistemic value is simplicity.

Simplicity is an extrinsic epistemic value for it can be truth-promoting, but only in combination with some other intrinsic epistemic value like accuracy, at least a sufficient degree of it. Extrinsically epistemic status saves its epistemic role without commitments to generality, because circumstances matter. In contrast, empirical accuracy is an intrinsic epistemic value, and also a robust one, "in the sense of being epistemic in almost any setting", while most other epistemic values Steel calls contextual because "their capacity to promote the attainment of truth depends on occurring within a specific set of circumstances" (2010, 20).[2] Similar to simplicity, Steel would be consistent to argue that speed is a contextual and extrinsic epistemic value because it can promote the attainment of truth, but that depends on the appropriate degree of accuracy involved. In both cases discussed earlier it is precisely such a value, for it has an epistemic role granted by an accompanying degree of accuracy. This role consists in avoiding the cost of suspended judgment, which is avoiding a situation that does not bring us closer to truth.

[2] Note that his use of "contextual" is not the same as Longino's in her distinction between constitutive and contextual values (Longino 1990).

Steel's account of epistemic is not contrasted with evaluative, social, historical, contingent, or contextual in Longino's sense (2010, 23), so it allows a broader scope of factors to count as extrinsically epistemic values, such as fundability or diversity of viewpoints, as long as they play a role in attaining the truth. This is why I contend that time-sensitivity might be considered as one of Steel's extrinsically epistemic values. In the two cases from the beginning in which time-sensitivity was introduced, it was motivated by non-epistemic considerations, but since it didn't compromise epistemic norms, even more, it promoted speed and therefore served an epistemic purpose of moving us temporally closer to truth, it was certainly acting extrinsically epistemic by promoting the attainment of truth in the given circumstances.

The problem with Steel's account is that it fails to discern between extrinsically epistemic values and non-epistemic values, especially when their influence is legitimate. The central aim of his account is to save the epistemic/non-epistemic distinction because of its usefulness in the argument from inductive risk. In order to do that, he develops "a principled basis for separating legitimate from illegitimate influences of non-epistemic values in scientific inference", (2010, 14) which states that "influences of non-epistemic values on scientific inferences are epistemically bad if and only if they impede or obstruct the attainment of truths." (2010, 15) In other words, influences of non-epistemic values are epistemically harmless if they don't impede or obstruct the truth. In fact, if they are not only harmless, but also beneficial in guiding us towards truth, as I claim they can be, we can call them extrinsically epistemic. Let us take a closer look.

Steel analyses two cases in which influence of non-epistemic values is welcome, to show how this is possible. The first case is precisely about speed – how long to wait or how much data to collect before accepting or rejecting a hypothesis, and the other is about judging some mistakes worse than others. I will limit this analysis to the first type of cases. We have already seen that favoring speed, i.e. not waiting and not collecting additional, more detailed data, can be epistemically beneficial. I see no reason to regard this case as non-epistemic-but-legitimately-influencing, when it fits perfectly well under the scope of extrinsically epistemic values. If the default position of speed is for Steel extrinsically epistemic, as I contend it is, then what is non-epistemic, for example in the expedited assessment case, is the protection of human health as a value that motivates expedited risk assessments in the first place. If it doesn't obstruct the attainment of truth, but often promotes it (we can't help people by pursuing untruthful and time-insensitive practices), why wouldn't we grant it an extrinsically epistemic status as well? There is no reason for separating the status of speed and the protection of human health in this particular case when the only criterion is their relation to the attainment of truth. After all, the circumstances matter. The protection of human health in these circumstances meets the condition of an extrinsically epistemic value. This becomes even clearer if we contrast it to fundability or diversity of viewpoints whose default position in Steel's account is extrinsically epistemic. There seems to be no problem in calling fundability and diversity of viewpoints non-epistemic-but-legitimately-influencing in *some* cases. There is no grounded difference between that status and an extrinsically epistemic status.

Steel's motivation is clear: he wants to save the argument from inductive risk which claims that non-epistemic values sometimes legitimately influence scientific research. And they do, but I claim that in those cases we can also call them extrinsically epistemic. They don't impede or obstruct the attainment of truth and they often point in the direction of truth as, for example, time-sensitive practices, speed of getting at true results that they promote, and the protection of human health and economic benefits that motivate these time-sensitive practices. Introducing the intrinsic/extrinsic distinction didn't save the distinction between epistemic and non-epistemic in the way Steel hoped it would. Now there is no proper scope for non-epistemic-but-legitimately-influencing, because extrinsically epistemic values have appropriated it, along with some of the values that used to be encountered on the lists of epistemic values, like simplicity and external consistency. Either there are only intrinsic epistemic values (namely, only empirical accuracy and internal consistency), and everything else is sometimes extrinsically epistemic (when it directs towards the truth in the given circumstances), otherwise it is non-epistemic because it doesn't have anything to do with the truth-seeking endeavor; or there are robust and intrinsic epistemic values and everything else is non-epistemic, but sometimes legitimately influencing scientific research. In any case, one side of the dichotomy has to be broadly construed, be it the epistemic or the non-epistemic side.

Steel endorsed a broad notion of epistemic which doesn't fall in line with the usual epistemic side of the dichotomies (internal-external, fact-value, direct-indirect, constitutive-contextual etc.), but is constrained only by the relation to the attainment of truth. The alternative would be to be rigid on the epistemic side and count only intrinsic

epistemic values as epistemic, and then carefully assess particular cases to allow for a legitimate influence of non-epistemic values in particular instances of research assessed on a case to case basis. Non-epistemic values would then have to be broadly construed to involve both simplicity and external consistency. In fact, we are left with particularism about what is epistemic and what is non-epistemic in specific cases of scientific research. I don't think that this is bad news, but it does show that Steel's distinction doesn't deliver on its promises.

More importantly, the notion of values, especially as broadly construed as Steel's, might be a misleading one in the first place. After all, not everything that we can talk about in this context is a *value*. As Justin Biddle puts it: "There are different factors that can fill the gap between 'insight' (i.e. logic, evidence, and epistemic values broadly construed) and decision making in science." (Biddle 2013, 132) I believe that time-sensitivity is a good example of such a factor.

5. Time-sensitivity

The debate in which the notion of time-sensitivity is introduced provides us with understanding of both its non-epistemic setting and its epistemic directedness. Sometimes we have social and pragmatic reasons to have the results quickly. Sometimes a scientist may want to have the results soon in order to move forward with her career or research, even if she is honestly dedicated to truth. Scientific work is embedded in time-frames: of funding, career stages, a lifetime, a generation or of several generations. Whatever the

reasons may be, we will want to assign a desired time-frame for achieving certain ends in sight, even when it comes to "building block" science. We do want to see *some* results at *some* time. The assigned value of the desired time-frame is the level of time-sensitivity, and it can and does affect how different values pertaining to research practices are balanced against each other, most obviously speed and accuracy of methods. The aim of attaining the truth doesn't only inform our methodological choices, it happens in time. It most certainly reflects an epistemic end, but is also motivated by all kinds of values and reasons. It would be misleading to call it a value, because it enters the picture as a judgment that has a say on how different values "hang" together. Even if the level of time-sensitivity is very low, it still is present.

For example, in basic science like gravitational wave physics, it takes a lot of time, computational power and extremely sensitive instruments to handle all the uncertainties related to the end-in-sight. Not long ago, the end-in-sight was the detection of gravitational waves. The time-sensitivity might have been estimated low at the beginning, especially since there are no immediate applications of the research; for now it has "only" yielded the benefit of better understanding of the universe and matter. In this case it was reasonable to expect decades of research without a robust result. However, with time passing by, the time-sensitivity of the detection attempts have grown nevertheless, because of huge cognitive and material investments which at some point require payoffs. Time-sensitivity motivates new procedures for error estimates, adding of computational power, and refinements of the instruments. Speeding up means coming up with new ways to get to the result, only in this context the tolerance for huge time-spans is higher. However, the

tolerance is also exhaustive if there is no measurable advancement. This will first be reflected in the shortage of funds and then in the shortage of researchers' interest.

This particular research succeeded: gravitational waves were first detected on September 14, 2015, after more than 50 years of research. However, there are no guaranties that every research will be as successful as that, and it especially won't be the case that 50 years will be an acceptable time-span for every research practice. In comparison, recent efforts around translation in biomedical sciences are in part a reaction to the fact that the average time-span between discovery and implementation of therapeutic practices, which has been estimated 17 years (Contopoulos et al. 2008, Morris et al. 2011), is considered way too long. Unsurprisingly, since the deliverances and applications of biomedical sciences are expected with much greater urgency than that of gravitational science. This judgment is so strong that it initiated a new model of biomedical research, namely translational science, dedicated to speeding up of the so called "bench to bedside" process. Time-sensitivity does have a saying on what the next step is and which values to prioritize in different research contexts.

The examples discussed in the beginning of the paper elucidate the fact that a certain degree of time-sensitivity is present in the context in which scientific research is done, in the uses it has, and problems it aims to address. The degree of time-sensitivity is implicitly or explicitly estimated and it has a bearing on the trade-offs between different values, such as speed and accuracy. As we have seen, simplicity and ease of use transitively address time-sensitivity by contributing to speed of methods and practices. A method can generate results faster in comparison to another, and those results can be more or less accurate, but

how much the setting of this activity is time-sensitive is a contextual and evaluative judgment that gives rise to concerns about efficiency and has a saying on how different methodological values are balanced against each other in particular instances of research. It doesn't fall exclusively under either epistemic or non-epistemic side of the dichotomy, it is rather informed by both: the aim of attainment of truth and the peculiarities of here and now. Highly time-sensitive issues favor expedited methods, in other words: higher the time-sensitivity, more valuable the speed.

6. Conclusion

In this paper I proposed an account of time-sensitivity, a notion introduced in Daniel Steel's comment (2016) on Elliott and McKaughan (2014). Time-sensitivity is a feature of problems to be solved in their particular contexts, a feature recognized by an implicit or explicit evaluative judgment about a desired or expected time-frame of having a result which gives rise to concerns about efficiency and influences methodological choices. I firstly pointed to speed as a value of research methods and practices that most specifically addresses time-sensitivity. Then I argued along the lines of Steel (2010, 2016) why speed ought to be considered an epistemic value, contrary to Elliott and McKaughan (2014). After that I tried to account for time-sensitivity by using Steel's distinction between extrinsically and intrinsically epistemic values (2010). I showed that his distinction fails to distinguish between extrinsically epistemic values and non-epistemic values, especially when their influence on research is legitimate.

References

Biddle, Justin (2013), "State of the field: Transient underdetermination and values in science", *Studies in History and Philosophy of Science* 44: 124-33.

Contopoulos-Ioannidis, Despina G., Alexiou, G. A., Gouvias, T. C., Ioannidis, J. P. A. 2008. "Life Cycle of Translational Research for Medical Interventions*", Science* 321(5894): 1298-99.

Cranor, Carl (1993), *Regulating Toxic Substances*. Oxford: Oxford University Press.

--- (1995), "The Social Benefits of Expedited Risk Assessments", Risk Analysis 15: 353-58.

Elliott, Kevin; McKaughan, Daniel, (2014), "Nonepistemic Values and the Multiple Goals of Science*", Philosophy of Science* 81(1): 1-21.

Longino, Helen (1990), *Science as social knowledge.* Princeton: Princeton University Press.

Morris, Zoë Slote, Wooding, S., Grant, J. 2011. "The answer is 17 years, what is the question: understanding time lags in translational research", Journal of the Royal Society of Medicine 104: 510-20.

Robertson, Morgan (2004), "The Neoliberalization of Ecosystem Services: Wetland Mitigation Banking and Problems in Environmental Governance", *Geoforum* 35: 361-73.

--- (2006), "The Nature That Capital Can See: Science, State, and Market in the Commodification of Ecosystem Services", *Environment and Planning D*, *Society and Space* 24: 367-87.

Steel, Daniel (2010), "Epistemic Values and the Argument from Inductive Risk", *Philosophy of Science* 77: 14-34.

--- (2016), "Accepting an Epistemically Inferior Alternative? A Comment on Elliott and McKaughan", *Philosophy of Science* 83: 606-12.

Word count: 4997

A statistical learning approach to a problem of induction

Kino Zhao
yutingz3@uci.edu

University of California, Irvine
Logic and Philosophy of Science

(Draft updated December 7, 2018)

Abstract

At its strongest, Hume's problem of induction denies the existence of any well justified assumptionless inductive inference rule. At the weakest, it challenges our ability to articulate and apply good inductive inference rules. This paper examines an analysis that is closer to the latter camp. It reviews one answer to this problem drawn from the VC theorem in statistical learning theory and argues for its inadequacy. In particular, I show that it cannot be computed, in general, whether we are in a situation where the Vapnik-Chervonenkis (VC) theorem can be applied for the purpose we want it to.

Hume's problem of induction can be analyzed in a number of different ways. At the strongest, it denies the existence of any well justified assumptionless inductive inference rule. At the weakest, it challenges our ability to articulate and apply good inductive inference rules. This paper examines an analysis that is closer to the latter camp. It reviews one answer to this problem drawing from a theorem in statistical learning theory and argues for its inadequacy.

The particular problem of induction discussed in this paper concerns what Norton (2014) calls a formal theory of induction, where "valid inductive inferences are distinguished by their conformity to universal templates" (p.673). In particular, I focus on the template that is often called *enumerative induction*. An inductive argument of this type takes observations made from a small and finite sample of cases to be indicative of features in a large and potentially infinite population. The two hundred observed swans are white, so all swans are white. Hume argues that the only reason we think a

rule like this works is because we have observed it to work in the past, resulting in a circular justification.

Nevertheless, this kind of inductive reasoning is vital to the advancement of a scientific understanding of nature. Most, if not all, of our knowledge about the world is acquired through the examination of only a limited part of the world. The scientific enterprise relies on the assumption that at least some of such inductive processes generate knowledge. With this assumption in place, a weak problem of induction asks whether we can reliably and justifiably differentiate the processes that do generate knowledge from the ones that do not. This paper discusses this weak problem of induction in the context of statistical learning theory.

Statistical learning theory is a form of supervised machine learning that has not received as much philosophical attention as it deserves. In a pioneering treatment of it, Harman and Kulkarni (2012) argue that one of the central results in statistical learning theory – the result on Vapnik-Chervonenkis (VC) dimensions – can be seen as providing a new kind of answer to a problem of induction by providing a principled way of deciding if a certain procedure of enumerative induction is reliable. The current paper aims to investigate the plausibility of their view further by connecting results about VC dimension in statistical learning with results about NIP models in the branch of logic called model theory. In particular, I argue that even if Harman and Kulkarni succeed in answering the problem of induction with the VC theorem, the problem of induction only resurfaces at a deeper level.

The paper is organized as follows: section 1 explains the relevant part of statistical learning theory, the VC theorem, and the philosophical lessons it bears. Section 2 introduces the formal connection between this theorem and model theory and proves the central theorem of this paper. Section 3 concludes with philosophical reflections about the results.

1 Statistical learning theory

The kind of problems that is relevant for our discussion of VC dimensions is often referred to as classification problems that are irreducibly stochastic. In a classification problem, each individual is designated by its k-many features such that it occupies somewhere along a k-dimensional feature space, χ. The goal is to use this information to classify potentially infinitely many such individuals into finitely many classes. To

give an example, consider making diagnoses of people according to their test results from the k tests they have taken. The algorithm we are looking for needs to condense the k-dimensional information matrix into a single diagnosis: sick or not. The algorithm can be seen as a function $f : \chi \to \{0, 1\}$, where 1 means sick and 0 means not. For reasons of simplicity, I will follow the common practice and only consider cases of binary classification.

By "irreducibly stochastic", I mean that the target function f cannot be solved analytically. This might be because the underlying process is itself stochastic – it is possible for two people with exact same measures on all tests to nevertheless differ in health condition – or because the measurements we take have ineliminable random errors. This means that even the best possible f will make some error, and so the fact that a hypothesis makes errors in its predictions does not in itself count against that hypothesis. Instead, a more reasonable goal to strive towards is to have a known, preferably tight, bound on the error rate of our chosen hypothesis.

What makes this form of statistical learning "supervised learning" is the fact that the error bound of a hypothesis is estimated using data points whose true classes are known. Throughout this paper, I will use D to denote such a dataset. D can have any cardinality, but the interesting cases are all such that D is of finite size. Recall that the feature (or attribute) space χ denotes the space of all possible individuals that D could have sampled, so that $D \subset \chi$. I understand a hypothesis to be a function $h : \chi \to \{0, 1\}$. A set of hypotheses $\mathcal{H}$ is a set composed of individual hypotheses. Usually, the hypotheses are grouped together because they share some common features, such as all being linear functions with real numbers as parameters. This observation will become more relevant later.

One obvious way of choosing a good hypothesis from $\mathcal{H}$ is to choose the one that performs the best on D. I will follow Harman and Kulkarni (2012) and call this method enumerative induction, for it bears some key similarities with Hume's description of the observation of swans. This method is inductive because it has the ampliative feature of assuming that the chosen hypothesis will keep performing well on individuals outside of D. The question we are interested in is: how do we know this? What justifies the claim that the hypothesis performs well on D will perform well outside of D too? The answer that will be examined in this section and throughout the rest of the paper is that we know this claim to be true when we are in a situation where $\mathcal{H}$ has finite VC dimension, and the VC-theorem justifies this claim.

To define the error rate of a hypothesis, recall the "ideal function" f mentioned in the introduction. Recall also that f classifies individuals from χ into $\{0, 1\}$, and f is imperfect. Nevertheless, since the process from χ to the classes is irreducibly stochastic, f is as good as we can hope for. Therefore, f will serve as our standard for the purpose of calculating the error rate of a hypothesis. Note that the hypotheses we are assessing are all from $\mathcal{H}$, our hypothesis set, but f need not be in $\mathcal{H}$.

Suppose D is of size N, and $x_1, \ldots, x_N \in D$. For each $h \in \mathcal{H}$ and $i \in [1, N]$, consider the random variable $X_i : \chi^N \to \{0, 1\}$ defined by

$$X_i\big(h(x_1, \ldots, x_N)\big) = \begin{cases} 1 & \text{if } h(x_i) \neq f(x_i), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Intuitively, $X_i = 1$ if the hypothesis we are evaluating, h, gives a different (and hence wrong) verdict on x_i than the target function f, and 0 otherwise. Assume $X_1, \ldots, X_N$ are independent, which is to say that making a mistake on one data point does not make it more or less likely for h to make a mistake on another one. This is typical if D is obtained through random sampling. Further assume $X_1, \ldots, X_N$ are identically distributed, which means that for any X_i and X_j in the sequence, $EX_i = EX_j$. This allows the error "rate" of h across multiple data points to be meaningfully computed.

Let $\overline{X} = \frac{1}{N}(\sum_{i=1}^{N} X_i)$, which is the measured mean error, and $\mu = E\overline{X}$, which is the expected mean error. I will follow Abu-Mostafa et al. (2012) in calling the former the *in-data error*, or E_{in}, and the latter *out-data error*, or E_{out}. To flesh out the relationship between these two values more clearly, we define

$$E_{in}(h) = \overline{X} = \frac{1}{N} \sum_{i=1}^{N} [\![h(\mathbf{x}_i) \neq f(\mathbf{x}_i)]\!] \tag{2}$$

$$E_{out}(h) = \mu = \mathbb{P}_N(h(\mathbf{x}) \neq f(\mathbf{x})) \tag{3}$$

Intuitively, the in-data error is the evidence we have about the performance of h, and the out-data error is the expectation that h will hold up to its performance. The amplification comes in when we claim that E_{out} is not very different from E_{in}. I will call the difference between E_{in} and E_{out} the *generalization error*.

For any single hypothesis, and for any error tolerance $\epsilon > 0$, Hoeffding (1963, p.16) proved a result called the *Hoeffding inequality* (see also Lin and Bai 2010, p. 70, and

Pons 2013, p. 205), which states that, under the assumption that the error rate for each data point is independent and identically distributed, we have (in the notations introduced above)

$$\mathbb{P}_N(|E_{in}(h) - E_{out}(h)| \geq \epsilon) \leq 2e^{-2\epsilon^2 N} \tag{4}$$

This inequation says that the probability of having a large generalization error in the assessment of a single hypothesis is bounded by $2e^{-2N\epsilon^2}$, which is a function of the size of the dataset, N, and the error tolerance ϵ.

Once we establish a bound in the case of a single hypothesis, we can get a similar bound for finitely many such hypotheses. The reason we cannot simply apply the Hoeffding inequality to our preferred hypothesis is that it requires us to pick a hypothesis before we compute its error rate from the data. This will not help us if we need to use data to do the picking. Instead, we need to make sure *any* hypothesis we pick out will have low enough generalization error, before we can trust the method (of enumerative induction) we use to pick.

Since we assume that the error rate of one hypothesis is independent of another, the probability of any of the finitely many hypotheses we are considering having a large generalization error is just going to be the union of the probability of each one of them does. In symbolic form, suppose there are $1 \leq M < \infty$ many hypotheses in $\mathcal{H}$, then we have

$$\mathbb{P}(\max_{h\in\mathcal{H}} |E_{in}(h) - E_{out}(h)| \geq \epsilon) = \mathbb{P}(\exists h \in \mathcal{H} |E_{in}(h) - E_{out}(h)| \geq \epsilon) \leq 2Me^{-2\epsilon^2 N} \tag{5}$$

While this bound may seem "loose", it serves our purpose when we have a reasonably small M or a reasonably large N.

This simple calculation becomes tricky, however, when $\mathcal{H}$ contains infinitely many hypotheses. If we replace M with infinity, then the upper bound stops being a bound, because $2Me^{-2\epsilon^2 N}$ grows to infinity as M does. This is where the VC dimension of $\mathcal{H}$ comes to play.

To understand the role of VC dimensions, define

$$\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \{(h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N)) \mid h \in \mathcal{H}\} \tag{6}$$

which is the set of all verdicts given by $\mathcal{H}$ on dataset D. If some hypotheses agree with each other on the classification of every data point, then their verdicts would be represented by the same tuple. This means that the cardinality of the set of verdicts

may be much smaller if $\mathcal{H}$ is very homogeneous. Moreover, different datasets of the same cardinality may elicit more or fewer different verdicts from $\mathcal{H}$. Define

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1,\cdots,\mathbf{x}_N \in \chi} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)| \tag{7}$$

as the max number of different verdicts $\mathcal{H}$ can generate from any dataset of cardinality N.

If all possible classifications of D have been represented in $\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)$, then we have $m_{\mathcal{H}}(N) = 2^N$. When this happens, we say that the hypothesis set $\mathcal{H}$ *shatters* the dataset D. Define the *VC dimension of* $\mathcal{H}$ to be the maximum N such that $m_{\mathcal{H}}(N) = 2^N$. In other words, it is the maximum number N such that there exists a dataset D of size N that is shattered by $\mathcal{H}$. If $m_{\mathcal{H}}(N) = 2^N$ holds for all N, then we say the VC dimension is infinite. Let's call a hypothesis set $\mathcal{H}$ VC-learnable if it has finite VC dimension.

Very roughly, the VC dimension of a hypothesis set tracks the maximum number of hypotheses that are still distinguishable from each other with respect to their verdicts on data. This means that, if we consider any more hypotheses, some of them will always agree with some others on all of the classifications they give to all possible data points, and so if one has low generalization error, the others will, too. The VC generalization bound is given as follows (Abu-Mostafa et al., 2012, p.53)

$$\mathbb{P}_N[\![\big(E_{out}(h) - E_{in}(h)\big) \leq \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}]\!] \geq 1 - \delta \tag{8}$$

where δ is the uncertainty tolerance. If $\mathcal{H}$ has an infinite VC dimension, then no such upper bound can be found. Notice that, holding everything else equal, increasing N brings the right-hand side down, which means that increasing data size allows us to make a better estimate of E_{out} with the same uncertainty tolerance. One can further show that

$$\lim_{N\to\infty} \mathbb{P}_N(\max_{h\in\mathcal{H}} |E_{in}(h) - E_{out}(h)| = 0) > 1 - \delta \tag{9}$$

for all $\delta > 0$. This means that, when $\mathcal{H}$ is either finite or has finite VC dimension, we can justifiably claim enumerative induction to be a reliable process that can pick out a good hypothesis from $\mathcal{H}$.

What makes this theorem especially powerful is not just that it shows how the error rates converge in the limit, but also that the convergence is uniform. What is

practically useful for statisticians is not so much that, if we have infinite data, we can figure out the true error rate of our hypothesis, but that, as soon as we know how many data points we have and the VC dimension of $\mathcal{H}$, we know precisely how confident we should be of our estimation of the error rate.

In what sense does this theorem answer a problem of induction? According to the analysis in Harman and Kulkarni (2012), this theorem defines precise conditions (i.e., ones where $\mathcal{H}$ has finite VC dimension) under which a particular inductive method (i.e., supervised learning in classification problems) is reliable. To the extent that we are concerned with the "easy" problem – the practical problem – of induction, the VC theorem does seem to provide a kind of answer we are looking for. In the next section, I challenge the applicability of this answer. In particular, I show that we can never know in general if we are in a situation where the above answer is applicable.

2 Finiteness of VC dimensions is uncomputable

A preliminary observation about the finiteness requirement is that we do not have a good grasp of what it means. What is the difference between these two sets of hypotheses such that one has finite VC dimension and the other does not? To put this point more concretely, we know that polynomial functions with arbitrarily high degrees have finite VC dimension, whereas the set of formulas with the sine function has infinite VC dimension. What is the difference between them? If we have a problem that can be reasonably formulated as polynomials or with a sine function, do we have good principled reasons why we should formulate it in one way rather than another?

Surprisingly, model theory in logic might help shed light on this question. It turns out that the concept of NIP theories corresponds to the class of hypothesis sets with finite VC dimensions. A theorem provably equivalent to the VC theorem was independently proved by the model theorist Shelah about these NIP theories and the corresponding NIP models. This connection was first recognized by Laskowski (1992). Interestingly, with the real numbers as their underlying domains, models with the usual plus and multiplication signs are NIP, whereas adding the sine curve makes them not NIP. This suggests that we can ask the same questions we would like to ask about our statistical hypothesis sets in model theory, which has a richer structure that is better understood independently.

In the previous section we discussed how the idea of "distinguishable hypotheses"

is important for the VC theorem. If a hypothesis set has finite VC dimension, we can think of it as having finitely many *distinguishable* hypotheses, even if it in fact has infinitely many. Intuitively speaking, if our dataset is "large enough" that not every combination of verdicts is representable with our hypotheses, then we can talk about which hypothesis is truly better than its competitors, as opposed to accidentally matching the specific data points. Having finite VC dimension ensures that there exist finite datasets that are "large enough". If a hypothesis set has finite VC dimension, let us call the set *VC-learnable*.

The corresponding concept in model theory relies on the same idea of distinguishability. Intuitively, if a formula is NIP – has the not-independent property – then there exists a natural number n such that no set larger than that number can be defined using this formula. A model is NIP just in case all of its formulas are (a formal definition is presented in Appendix A; for more formal details, see Simon, 2015).

We can then treat each hypothesis set as a formula defined on some domain. Laskowski (1992) shows that a hypothesis set is VC-learnable just in case the corresponding formula is NIP. What makes this correspondence especially useful is that model theorists have devoted a lot of efforts into determining which model is NIP. Once we know of a model that it's NIP, we also know that any hypothesis sets formulated using the language and domain of this model are VC-learnable.

For example, there is a group of models called *o-minimal*, which roughly means that all the definable subsets of the domain are finite unions of simple topological shapes like intervals and boxes. It suffices for our purposes to note that all o-minimal models are NIP (van den Dries, 1998, p. 90). As it happens, the real numbers with just addition and multiplication are o-minimal (van den Dries, 1998, p. 37). This means that any hypothesis set consisted of addition, multiplication, and the real numbers are going to have finite VC dimension. Similarly, the real numbers with addition, multiplication, and exponentiation is also o-minimal (Wilkie, 1996). This means that all sets of polynomials are VC-learnable.

As alluded to already, the real numbers with the sine function added are not NIP. This is roughly because, with the sine function, we can define copies of the integers using the set $\{x \in \mathbb{R} : \sin(x) = 0\}$, which allows us to define all of second-order arithmetic, and second-order arithmetic allows coding of arbitrary finite sets. As expected, this is reflected in statistical learning theory by the fact that the set of sine functions has infinite VC dimension, and so is not VC-learnable.

2. FINITENESS OF VC DIMENSIONS IS UNCOMPUTABLE

Another important observation from model theoretic investigations on NIP theory is that there seem to be no easy test for when an expansion of the real numbers is NIP. Although the relationship between the NIP property and properties like o-minimal and stable (a set of structures that are not o-minimal but are NIP) is well-researched and understood, there is no uniform way of telling where exactly a model lies (see, e.g., Miller, 2005[1]).

The statistical learning community echoes this difficulty with the observation that "it is not possible to obtain the analytic estimates of the VC dimension in most cases" (Shao et al., 2000; also see Vapnik et al., 1994). Recall that the VC dimension decides how big a dataset is "big enough". If the view is that enumerative induction is a reliable method when we are confident (i.e., low δ) that its assessment of hypotheses generalizes (i.e., low ϵ) and the VC theorem is supposed to guarantee this, then our inability to analytically solve the VC dimension of a given hypothesis set seems deeply handicapping.

To make the matter worse, it turns out that even knowing when we do have finite VC dimension is not a straightforward task, as witnessed by the following theorem, whose proof is given in Appendix A

Theorem 1. *The set $\{\varphi(x,y) : \varphi(x,y)$ is $NIP\}$, where $\varphi(x,y)$ is formulated in the language of arithmetic with addition and multiplication, is not decidable. In particular, this set computes $\emptyset^{(2)}$, the second Turing jump of the empty set.*

What this theorem tells us is that, in general, there is no effective procedure we can follow that can tell us, for any 2-place formula $\varphi(x,y)$, if it's NIP. With Laskowski's result, this means that we cannot compute, in general, if a given hypothesis set is VC-learnable either.

The specific way in which the set of all NIP formulas is uncomputable is significant also. For some time now, philosophers who study knowledge and learning from a formal perspective have placed a lot of emphasis on learning in the limit. Kelly (1996, p.52), for example, argues that the concept of knowledge (as opposed to, say, mere belief) implies that the method of generating such beliefs is stable in the limit. He then argues that the best way to formalize the notion of "stability in the limit" is to understand it as computable in the limit. Relatedly, a venerable tradition of formal learning

[1]Technically, Miller is interested in dichotomy theorems which establish either that an expansion of the reals is o-minimal or that it defines second-order arithmetic. As mentioned before, the former suffices for being NIP, and the latter suffices for being not NIP.

theory following Gold (1967) has explored extensively the conditions under which a noncomputable sequence may or may not be approximated by a computable sequence making only finitely many mistakes (cf. Osherson et al., 1986; Jain et al., 1999). From this perspective, it seems we might still be able to claim knowledge of what is or isn't knowable if we can compute the set of NIP formulas in the limit. Unfortunately, this latter task cannot be accomplished. This is because that, in order for a sequence to be approximable in the limit by another sequence, it cannot be harder than the first Turing jump of the sequence used to approximate it (Soare, 1987, p.57; see also Kelly, 1996, p.280). This means that something that is at least as hard as the second Turing jump cannot be approximated by a computable sequence.

To recapitulate the dialectic so far: an easy problem of induction asks us to identify and then justify the conditions under which a given ampliative method is reliable. The VC theorem gives one answer: supervised statistical learning from data is reliable just in case the hypothesis set has finite VC dimension. However, it turns out that we cannot, in general, decide if a hypothesis set is VC-learnable.

Can we judge our $\mathcal{H}$ on a case-by-case basis? Once we fix an $\mathcal{H}$, we can usually tell if it has finite VC dimension, and we can develop methods of empirically estimating its VC dimension using multiple datasets with varying sizes. However, this seems to just push the same problem to a deeper level. The problem that a method "sometimes is reliable, sometimes isn't", is solved by specifying a condition under which it always is reliable. But the problem that the condition "sometimes occurs, sometimes doesn't" seems to have no simple solution. In fact, the above theorem says that the latter problem has no solution.

3 Conclusion

A reasonable conclusion to draw from the discussions we've had so far, I think, is that the VC theorem still does not give us the kind of robust reliability we need to answer a question with some scope of philosophical generality. As is typical of answers people give to problems of induction, as soon as a rule is formulated, a question arises concerning its applicability. Similarly, what started out as a concern over the robustness of the method of enumerative induction turns into a concern over the robustness of the identifiable condition (i.e., the VC-learnable condition) under which enumerative induction is justified to be reliable.

A related question concerns the distinction, if there is one, between the cases where $\mathcal{H}$ has infinite VC dimension and cases where it has a VC dimension so large that it's impractical for us to make use of it. There is a sense in which the case of an infinite VC dimension fails *in principle*, whereas the case of a very large VC dimension only fails in *practice*. However, it is often impossible to analytically solve the VC dimension of a hypothesis set even if we do know that it's VC-learnable. Together with the result that we cannot test if a case is VC-learnable *in principle*, it seems to suggest that any information we might gain from the distinction between failing in principle and failing in practice will not be very informative, since we often can't tell which case we are in.

The philosophical difficulties discussed above raise an interesting question of how the practitioners view the same obstacle. Perhaps the way out is to accept a 'piecemeal' solution after all. It seems that when the VC dimension is small, we can often know both that it is finite, and that it is small. Theorists have also developed ways of estimating VC dimension using multiple datasets (see, e.g., Vapnik et al., 1994 and Shao et al., 2000). It seems that, as it often happens, philosophical problems are much more manageable when we do not look for principled solutions.

Acknowledgement

I would like to express my gratitude towards Sean Walsh for his supervision, as well as towards the participants in the 2016 Logic Seminar and attendees of the Society for Exact Philosophy 2017 meeting for their valuable feedback and discussion.

Appendix A

This appendix presents the proof of Theorem 1. I will follow the definition of NIP formulas given by Simon (2015) as follows (with notations changed to match preceding text)

> Let $\varphi(x;y)$ be a partitioned formula. We say that a set A of $|x|$-tuples is *shattered* by $\varphi(x;y)$ if we can find a family $(b_I : I \subseteq A)$ of $|y|$-tuples such that
>
> $$M \models \varphi(a;b_I) \iff a \in I, \quad \text{for all } a \in A$$

A formula $\varphi(x;y)$ is NIP if no infinite set of $|x|$-tuples is shattered by it.

3. CONCLUSION

Following notations from Soare (1987), let W_e to be the domain of the e-th partial recursive function and $Fin = \{e : W_e < \omega\}$.

Lemma Given e, define the following formula in the language of arithmetic

$$\begin{aligned}\theta_e(x,y) =& \exists l > x \; \exists \text{ enumeration } c_1, \dots c_{2^l}, \text{ first } 2^l \text{ elements of } W_e \\ & \wedge \exists |\sigma| = l \text{ with } y = c_\sigma \wedge \sigma(x) = 1\end{aligned}$$

Then $e \in Fin$ iff θ_e is NIP.

Proof. ($\Rightarrow$) Suppose $e \in Fin$. The claim is: there is finite number N such that $|W_e| \leq 2^N$, and for all n, if a set A with cardinality n is shattered by θ_e, then $n \leq N$.

In particular, we show that the claim holds for N being the size of W_e. For the sake of contradiction, suppose there is A, with size n, shattered by θ_e, and $n > N$.

Let $A = \{a_1, \dots, a_n\}$, $\{b_I : I \subset \{a_1, \dots, a_n\}\}$, such that $\theta_e(a_i, b_I)$ iff $a_i \in I$.

Without loss of generality, let $a_n \geq n - 1$, and $I = \{a_n\}$. Then $a_n \in I$, and $\theta_e(a_n, b_I)$. This means that $\exists l > a_n \geq n - 1$ with the first 2^l many elements of W_e enumerated. Recall that the reductio hypothesis states $n > N$. This means that $|W_e| \geq 2^l > 2^{n-1} \geq 2^N$. This contradicts the original assumption that $|W_e| \leq 2^N$.

($\Leftarrow$) To show the contrapositive of this direction, suppose $e \notin Fin$, $|W_e| = \omega$. The claim is: θ_e is IP. Namely, $\forall N \; \exists n \geq N$, with some set A of cardinality n that is shattered by θ_e.

Take an arbitrary $n \geq N$. Let $A = \{0, \dots, n-1\}$. Let b_σ's be the first 2^n elements of W_e, as σ ranges over finite strings of length n. Since σ is a string, we say $a \in \sigma \Leftrightarrow \sigma(a) = 1$.

We need to show that $\theta_e(a, b_\sigma) \Leftrightarrow \sigma(a) = 1$.

The left to right direction is trivial, since it is part of $\theta_e(a, b_\sigma)$ to state that $\sigma(a) = 1$.

To show the right to left direction, note that since $|W_e| = \omega$, there definitely exists an initial segment of 2^n many elements of W_e, and $n > a$ for all $a \in A$. This satisfies the first conjunct. To satisfy the second conjunct of θ_e, recall that we defined our enumeration to be such that $|\sigma| = n$ with σ being identified with every number $\leq 2^n$. This means that an enumeration of $c_1 \dots c_{2^n}$ includes all c_σ with $|\sigma| = n$. Define $b_\sigma = c_\sigma$, and we are guaranteed that b_σ is in the enumeration, and $|\sigma| = n$. Finally, the last conjunct of θ_e is satisfied by supposition.

□

Theorem. *The set* $\{\varphi(x,y) : \varphi(x,y)$ *is* $NIP\}$*, where* $\varphi(x,y)$ *is formulated in the language of arithmetic with addition and multiplication, is not decidable. In particular, this set computes* $\emptyset^{(2)}$*, the second Turing jump of the empty set.*

Proof. Suppose not, then for any formula $\varphi(x,y)$, we can decide if it's NIP. This means that, for any e, we can decide if $\theta_e(x,y)$ as defined in the lemma above is NIP. By lemma, $\theta_e(x,y)$ is NIP just in case $e \in Fin$. If we could decide the former, we would be able to decide the set Fin. But by Soare (1987, p.66, Theorem 3.2), Fin is Σ_2-complete, and so computes $\emptyset^{(2)}$, the second Turing jump of the empty set, and hence is not computable. □

Bibliography

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook Singapore.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Harman, G. and Kulkarni, S. (2012). *Reliable reasoning: Induction and statistical learning theory*. MIT Press.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Jain, S., Osherson, D. N., Royer, J., and Sharma, A. (1999). *Systems that learn: an introduction to learning theory*. MIT press.

Kelly, K. T. (1996). *The logic of reliable inquiry*. Oxford University Press.

Laskowski, M. C. (1992). Vapnik-Chervonenkis classes of definable sets. *Journal of the London Mathematical Society*, 45(2):377–384.

Lin, Z. and Bai, Z. (2010). *Probability inequalities*. Science Press Beijing, Beijing; Springer, Heidelberg.

Miller, C. (2005). Tameness in Expansions of the Real Field. In *Logic Colloquium '01*, volume 20 of *Lecture Notes in Logic*, pages 281–316. Associaton for Symbolic Logic, Urbana, IL.

Norton, J. D. (2014). A material dissolution of the problem of induction. *Synthese*, 191(4):671–690.

Osherson, D. N., Stob, M., and Weinstein, S. (1986). *Systems that learn: An introduction to learning theory for cognitive and computer scientists.* The MIT Press.

Pons, O. (2013). *Inequalities in analysis and probability*. World Scientific.

Shao, X., Cherkassky, V., and Li, W. (2000). Measuring the VC-dimension using optimized experimental design. *Neural computation*, 12(8):1969–1986.

Simon, P. (2015). *A Guide to NIP Theories*. Lecture Notes in Logic. Cambridge University Press, Cambridge.

Soare, R. I. (1987). *Recursively Enumerable Sets and Degrees*. Perspectives in Mathematical Logic. Springer, Berlin.

van den Dries, L. (1998). *Tame Topology and O-Minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge.

Vapnik, V., Levin, E., and Le Cun, Y. (1994). Measuring the VC-dimension of a learning machine. *Neural computation*, 6(5):851–876.

Wilkie, A. J. (1996). Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association

Seattle, WA; 1-4 November 2018

Version: 31 October 2018

PhilSci A·R·C·H·I·V·E

Table of Contents

Mixed-Effects Modeling and Non-Reductive Explanation

(4975 words)

Abstract: This essay considers a mixed-effects modeling practice and its implications for the philosophical debate surrounding reductive explanation. Mixed-effects modeling is a species of the multilevel modeling practice, where a single model incorporates simultaneously two (or even more) levels of explanatory variables to explain a phenomenon of interest. I argue that this practice makes the position of explanatory reductionism held by many philosophers untenable, because it violates two central tenets of explanatory reductionism: single level preference and lower-level obsession.

1. Introduction

Explanatory reductionism is the position which holds that, given a relatively higher-level phenomenon (or state, event, process, etc.), it can be reductively explained by a relatively lower-level feature (Kaiser 2015, 97; see also Sarkar 1998; Weber 2005; Rosenberg 2006; Waters 2008).[1] Though philosophers tend to have slightly different conceptions of the position, two central tenets of the position can still be extracted:[2]

> **Single level preference:** a phenomenon of interest can be fully explained by invoking features that reside at a single, well-defined level of analysis (e.g., molecular level in biology).

[1] According to Sarkar (1998), explanatory reduction is an epistemological thesis which is distinguished from constitutive (ontological) and theory reductionism theses. Kaiser further distinguishes two sub-types of explanatory reduction: (a) "a relation between a higher-level explanation and a lower-level explanation of the same phenomenon" (2015, 97); (b) individual explanations, i.e., given a relatively higher-level phenomenon, it can be reductively explained by a relatively lower-level feature (*Ibid*., 97). This essay will focus on the second sub-type. Besides, when referring to levels I mean either hierarchical organization such as universities, faculties, departments etc., or functional organization such as organs, tissues, cells etc. When referring to scales I mean spatial or temporal scaling where levels are not so clearly delimited.

[2] Similar summary of the position can be found in Sober (1999).

> **Lower-level obsession:** lower-level features always provide the most significant and detailed explanation of the phenomenon in question, so a lower-level explanation is always better than a higher-level explanation.

Philosophers sometimes express these two tenets explicitly in their work. For example, Alex Rosenberg holds that "[…] there is a full and complete explanation of every biological fact, state, event, process, trend, or generalization, and that this explanation will cite only the interaction of macromolecules to provide this explanation" (Rosenberg 2006, 12). Marcel Weber expresses a similar idea in his explanatory hegemony thesis, according to which it's always some lower-level physicochemical laws (or principles) that ultimately do the explanatory work in experimental biology (Weber 2005, 18-50). John Bickle attempts to motivate a 'ruthless' reduction of psychological phenomena (e.g., memory) to the molecular level (Bickle 2003).

However, many philosophers have questioned the plausibility of the position on the basis of scientific practice (Hull 1972; Craver 2007; Bechtel 2010; Brigandt 2010; Hüttemann and Love 2011; Kaiser 2015). To counter that position, some authors have pointed to the relevance of an important practice that has not received sufficient attention before: multiscale or multilevel modeling or sometimes called integrative modeling approach, where a set of distinct models ranging over multiple levels or scales—including the macro-phenomenon level/scale—are involved in explaining a (often complex) phenomenon of interest

(Mitchell 2003, 2009; Craver 2007; Brigandt 2010, 2013a, 2013b; Knuuttila 2011; Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017). Often these models work together by providing diverse constraints on the potential space of representation (Knuuttila and Loettgers 2010; Knuuttila 2011; Green 2013).

This multilevel modeling surely casts some doubt on explanatory reductionism, for it seems unclear what reductively explains what—all those facts in the set of models ranging over different levels/scales are involved in doing some explanatory work. However, there is a species of multilevel modeling that has slipped away from most philosophers' sights: mixed-effects modeling (MEM hereafter)—also called multilevel regression modeling, hierarchical linear modeling, etc.—in which a single model incorporating simultaneously two (or even more) levels of variables is used to explain a phenomenon. For a mixed-effects model to explain, features of the so-called reducing and reduced levels must be simultaneously incorporated into the model, that is, they must go hand in hand.

MEM deserves special attention because it sheds new light on the reductionism-antireductionism debate by showing that (a) a mixed-effects model violating the two central tenets of explanatory reductionism can provide successful explanation, and (b) a single mixed-effects model without integrating with other epistemic means can also provide such successful explanation. Therefore, MEM first further challenges the explanatory reductionist position, and

second offers a novel perspective bolstering the multilevel/multiscale integrative approach discussed by many philosophers.

The essay proceeds as follows. Section 2 discusses the challenges faced by the traditional single-level modeling approach, and examines the reasons why the MEM approach is preferable in dealing with these challenges. Section 3 describes a MEM practice using a concrete model. Section 4 elaborates on the implications of MEM for the explanatory reductionism debate. Finally, Section 5 considers potential objections to my viewpoint.

2. Challenges to Reductive Explanatory Strategies

In many fields (e.g., biological, social and behavioral sciences) scientists find that the data collected show an intrinsically hierarchical or nested feature. Consider a simple example: we might be interested in examining relationships between students' achievement at school (A hereafter) and the time they invest in studying (T).[3] In conducting such a research, we might collect data from different classes (say 5 classes in total), with each class providing the same number of samples (say 10 students in each class). The data collected among classes might be taken for granted to be independent. Then we may use certain traditional statistical techniques such as ordinary least-squares (OLS) to analyze the data and build a linear relationship between A and T.

[3] For scientific studies of this kind, see Schagen (1990), Wang and Hsieh (2012), and Maxwell et al. (2017).

However, this single-level reductive analysis can lead to misleading results, because it ignores the possibility that students within a class may be more similar to each other in important aspects than students from different classes. In other words, each group (class) may have its own features relevant to the relationship between A and T that the other groups lack. Hence, the data collected from the students are in fact not independent, i.e., the subjects are not randomly sampled, because the individuals (students) are clustered within groups (classes). In technical terms, we say our analysis may fall prey to the *atomistic fallacy* where we base our analysis solely on the individual level—i.e., we reduce all the group-level features to the individuals. Therefore, traditional OLS techniques such as multiple regression cannot be employed in this context, because the case under consideration violates a fundamental assumption of these techniques: the independence of observations (Nezlek 2008, 843).

Conversely, we may face the same problem the other way around if we fail to consider the inherently nested nature of the data. Consider the student-achievement-at-school case again. We may observe that in classes where the time of study invested by students is very high, the achievements of the students are also very high. Given such an observation, we may reason that students who invest a lot of time in studying would be more likely to get higher achievements at school. However, this inference commits the *ecological fallacy*, because it attributes the relationship observed at the group-level to the individual-level (Freedman 1999). The individuals may exhibit within-group differences that the single group-level analysis fails to capture. In technical terms, this inference flaws

because it reduces the variability in achievement at the individual-level to a group-level variable, and the subsequent analysis is solely based on group's mean achievement results (Heck and Thomas 2015, 3). Again, traditional statistical techniques such as multiple regression cannot be employed in this context.

In sum, a single-level modeling approach that disrespects the multilevel data structure can commit either an atomistic or an ecological fallacy. Confronted with these problems, one response is to 'tailor' the traditional statistical techniques by, e.g., adding an effect variable to the model which indicates the grouping of the individuals. However, many have argued that this approach is unpromising because it may give rise to enormous new problems (Luke 2004; Nezlek 2008; Heck and Thomas 2015). Alternatively, scientists have developed a new framework that takes the multilevel data structure into full consideration, i.e., the MEM approach, to which we now turn.

3. Case Study: A Mixed-Effects Model

Depending on different conceptual and methodological roots we have two broad categories of MEM approaches: the multilevel regression approach and the structural equation modeling approach. The former usually focuses on direct effects of predictor variables on (typically) a single dependent variable, while the latter usually involves latent variables defined by observed indicators (for details see Heck and Thomas 2015). For the purpose of this essay's arguments, I will concentrate on the first kind.

Consider the student-achievement-at-school example again. Since students are typically clustered in different classes, a student's achievement at school may be both influenced by her own features (e.g., time invested in studying) and her class's features (e.g., size of the class). Hence here comes two levels of analysis: the individual-level (level-1) and the group-level (level-2), and individuals ($i=1,2,\ldots,N$) are clustered in level-2 groups ($j=1,2,\ldots,n$).[4] Now suppose that students' achievements at school are represented as scores they get in the exam. The effect of time invested in studying on scores can be described as follows:

$$Y_{ij}=\beta_{0j}+\beta_{1j}X_{ij}+\varepsilon_{ij} \qquad (1)$$

where Y_{ij} refers to the score of individual i in the jth group, β_{0j} is a level-1 intercept representing the mean of scores for the jth group, β_{1j} a level-1 slope (i.e., different effects of study time on scores) for the predictor variable X_{ij}, and the residual component (i.e., an error term) ε_{ij} the deviation of individual i's score from the level-2 mean in the jth group. Equation (1) looks like a multiple regression model; however, the subscript j reveals that there is a group-level incorporated in the model. It can also be seen from this equation that both the intercept β_{0j} and slope β_{1j} can vary across the level-2 units, that is, different groups can have different intercepts and slopes.

[4] Note that, for instructive purposes, our case involves only two levels; however, the MEM approach can in principle be extended to many more levels.

The most remarkable thing of MEM is that we treat both the intercept and slope at level-1 as dependent variables (i.e., outcomes) of level-2 predictor variables. So here we write the following equations expressing the relationships between the level-1 parameters and level-2 predictors:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad (2)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \quad (3)$$

where β_{0j} refers to the level-1 intercept in level-2 unit j, γ_{00} denotes the mean value of the level-1 intercept, controlling for the level-2 predictor W_j, γ_{01} the slope for the level-2 variable W_j, and u_{0j} the error (i.e., the random variability) for unit j. Also, β_{1j} refers to the level-1 slope in level-2 unit j, γ_{10} the mean value of the level-1 slope controlling for the level-2 predictor W_j, γ_{11} the effect of the level-2 predictor W_j, and u_{1j} the error for unit j.

Equations (2) and (3) have specific meanings and purposes. They express how the level-1 parameters, i.e., intercept or slope, are functions of level-2 predictors and variability. They aim to explain variations in the randomly varying intercepts or slopes by adding one (or more) group-level predictor to the model. These expressions are based on the idea that the group-level characteristics such as group size may impact the strength of the within-group effect of study time on

scores. This kind of effect is called a *cross-level interaction* for it involves the impact of variables at one level of a data hierarchy on relationships at another level. We will discuss this in detail in the next section.

Now we combine equations (1), (2) and (3) by substituting the level-2 parts of the model into the level-1 equation. We finally obtain the following equation:

$$Y_{ij} = \left[\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + \gamma_{11} X_{ij} W_j\right] + \left[u_{1j} X_{ij} + u_{0j} + \varepsilon_{ij}\right] \qquad (4)$$

This equation can be simply understood that Y_{ij} is made up of two components: the fixed-effect part expressed by the first four terms and the random-effect part expressed by the last three terms. Note that the term $\gamma_{11} X_{ij} W_j$ denotes a cross-level interaction between level-1 and level-2 variables, which is defined as the impact of a level-2 variable on the relationship between a level-1 predictor and the outcome Y_{ij}. We have 7 parameters to estimate in (4), they are four fixed effects: intercept, within-group predictor, between-group predictor and cross-level interaction, two random effects: the randomly varying intercept and slope, and a level-1 residual.

Now a mixed-effects model has been built, and the next step is to estimate the parameters of the model. However, we will skip this step and turn to explore the philosophical implications of the modeling practice relevant to the explanatory reductionism debate.

4. Implications for the Explanatory Reductionism Debate

Looking closely into the MEM practice, we find that a couple of important philosophical implications for the explanatory reductionism debate can be drawn.

4.1. All levels are indispensable

The first, and most obvious, feature of MEM is that it routinely involves many levels of analysis in a single model, and all these levels are indispensable to the model in the sense that no level can be reduced to or replaced by the other levels. These levels consist of both the so-called reducing level in the reductionist's terminology, typically a lower-level that attempts to reduce another level, and the reduced level, typically a higher-level to be reduced by the reducing level. In our student-achievement-at-school case, for example, a reductionist may state that the group-level will be regarded as the reduced level whereas the student-level as the reducing level.

The indispensability of each level in the model can be understood in two related ways. First, due to the nested nature of data, only when we incorporate different levels of analyses to the model can we avoid either the atomistic or ecological fallacy discussed in Section 2. As discussed in the student-achievement-at-school example where students are clustered in different classes (in the manner that students from the same class may be more similar to each other in important aspects than students from different classes), reducing all the analyses to the level of individual students can simply miss the important

information associated with group-level features and thus lead to misleading results. Although it's true that the problem might be partially mitigated by tailoring traditional single-level analytical techniques such as multiple regression, it's also true that this somewhat ad hoc maneuver can simply bring about various new vexing and recalcitrant issues (Luke 2004; Nezlek 2008; Heck and Thomas 2015).

Second, the problem can also be viewed from the perspective of identifying explanatory variables. In building a mixed-effects model, the main consideration is often to find a couple of variables that may play the role of explaining the pattern or phenomenon observed in the data. Here a modeler must be clear about how to assign explanatory variables, for instance, she must consider if there are different levels of analyses and, if so, which explanatory variables should be assigned to what levels, and so on. These considerations may come before her model building because of background knowledge, which paves the way for her to develop a conceptual framework for investigating the problem of interest. However, without such a clear and rigorous consideration of identifying and assigning multilevel explanatory variables, an analysis can flaw simply because it confounds variables at different levels.

Respecting the multilevel nature of explanatory variables has another advantage: "Through examining the variation in outcomes that exists at different levels of the data hierarchy, we can develop more refined theories about how explanatory variables at each level contribute to variation in outcomes" (Heck and Thomas 2015, 33). In other words, in respecting the multilevel nature of

explanatory variables, we get a clear idea of how, and to what degrees, explanatory variables at different levels contribute to variation in outcomes. If these variables do contribute to variation in outcomes, as it always happens in MEM, then the situation suggests an image of *explanatory indispensability*: all the explanatory variables at different levels are indispensable to explaining the pattern or phenomenon of interest.

Given these considerations, therefore, one implication for the explanatory reductionism debate becomes clear: it isn't always the case that, given a relatively higher-level phenomenon it can be reductively explained by a relatively lower-level feature. Rather, in cases where the data show a nested structure or, put differently, the phenomenon suggests multilevel explanatory variables, we routinely combine the higher-level with the lower-level in a single (explanatory) model. As a result, one fundamental tenet of explanatory reductionism is violated: single level preference.

4.2. Interactions between levels

Another crucial feature of multilevel modeling is its emphasis on *a cross-level interaction*, which is defined as

> "The potential effects variables at one level of a data hierarchy have on relationships at another level […]. Hence, the presence of a cross-level interaction implies that the magnitude of a relationship observed within

> groups is dependent on contextual or organizational features defined by higher-level units". (Heck and Thomas 2015, 42-43)

Remember that there is a term $\gamma_{11} X_{ij} W_j$ in our mixed-effects model discussed in Section 3, which indicates the cross-level interaction between the group-level and the individual-level. More specifically, this term can be best construed as the impact of a group-level variable, e.g., group size, upon the individual-level relationship between a predictor, e.g., study time, and the outcome, e.g., students' scores.

The cross-level interaction points to the plain fact that an organization or a system can somehow influence its members or components by constraining how they behave within the organization or system. This doesn't necessarily imply top-down causation (Section 5.3 will turn back to this point). Within the context of scientific explanation, however, it does imply that it isn't simply that characteristics at different levels separately contribute to variation in outcomes, but rather that they interact in producing variation in outcomes. In other words, the pattern or phenomenon to be explained can be understood as generated by the interaction between explanatory variables at different levels. Therefore, to properly explain the phenomenon of interest, we need not only have a clear idea of how to assign explanatory variables to different levels but also an unequivocal conception of whether these explanatory variables may interact.

Different models can be built depending on different considerations of the cross-level interaction. To see this, consider the student-achievement-at-school

example again. In some experiment setting we may assume that there was no cross-level interaction between group-level characteristics and the individual-level relationship (between study time and scores). In such a situation, we kept the effect of individual study time on scores the same across different classes, i.e., we kept the slope constant across classes. In the meanwhile, we treated another group-level variable (i.e., intercept) as varying across classes, i.e., different classes have different average scores. So, this is a case where we have a clear idea of how to assign explanatory variables but no consideration of the cross-level interaction. Nonetheless, in a different experiment setting we may assume that there existed cross-level interaction, and hence the effect of individual study time on scores can no longer be kept constant across different classes. At the same time, we treated another group-level variable (i.e., intercept) as varying across classes. Hence, this is a case where we have both a clear idea of how to assign explanatory variables and a consideration of the cross-level interaction. Corresponding to these two different scenarios, two different mixed-effects models can be built, as shown below:

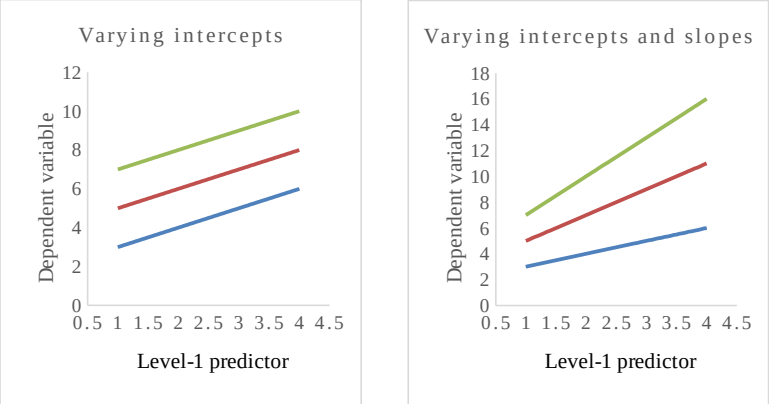


Figure 1. Two different models showing varying intercepts or varying intercepts and slopes, respectively. Three lines represent three classes. This figure is adapted from Luke (2004, 12).

Given such a cross-level interaction, therefore, the explanatory reductionist position has been further challenged. This is because any reductive explanation that privileges one level of analysis—usually the lower-level—over the others falls short of capturing this kind of interaction between levels. If they fail to do so, then they are missing important terms relevant to explaining the phenomenon of interest. As a consequence, a mixed-effects model involving interactions between levels simultaneously violates the two fundamental pillars of explanatory reductionism: first, it violates single level preference because it involves multilevel explanatory variables in explaining phenomena, and second, it violates lower-level obsession because it privileges no levels—all levels are interactively engaged in producing outcomes.

5. Potential Objections

This section considers two potential objections.

5.1. In-principle argument

One argument that resurfaces all the time in the reductionism-versus-antireductionism debate is the in-principle argument, the core of which is that even if reductive explanations in a field of study are not available for the time being, it doesn't follow that we won't obtain them someday (e.g., Sober 1999; Rosenberg 2006). Therefore, according to some reductionists, the gap between current-science and future-science is simply a matter of time, for advancement in techniques, experimentation and data collecting can surely fill in the gap.

However, I think the argument flaws. To begin with, advancement in techniques, experimentation and data collecting isn't always followed by reductive explanations. For example, in our MEM discussed in Section 3, even if the data about the individual-level is available and sufficiently detailed, it isn't the case that we explain the phenomenon of interest in terms of the data from the individual-level alone. Consider another example: in dealing with problems associated with complex systems in systems biology, even though large-scale experimentation (e.g., via computational simulation) can be conducted and high throughput data arranging over multiple scales/levels can be collected, a bottom-up reductive approach must be integrated with a top-down perspective so as to

produce useful explanations or predictions (Green 2013; Green and Batterman 2017; Gross and Green 2017).

Nevertheless, reductionists may reply that the situations presented above only constitute an in-practice impediment, for it doesn't undermine the *possibility* that lower-level reductive explanations, typically provided by some form of 'final science', will be available someday. Let us dwell on the notion of possibility a bit longer. The possibility here may be construed as a *logical possibility* (Green and Batterman 2017, 21; see also Batterman 2017). Nonetheless, if it's merely logically possible that there will be some final science providing only reductive explanations, then nothing can exclude another logical possibility that there will be some 'mixed-science' providing only multilevel explanations. After all, how can we decide which logical possibility is more possible (or logically more possible)? I doubt that logic alone could provide anything useful in justifying which possibility is more possible, and that appealing to logical possibility could offer anything insightful in helping us understand how science proceeds. As Batterman puts, "Appeals to the possibility of *in principle* derivations rarely, if ever, come with even the slightest suggestion about how the derivations are supposed to go" (2017, 12; author's emphasis).

Another interpretation of possibility may be associated with real possibilities, referring to the actual cases of reductive explanations happening in science. Unfortunately, I don't think the real scenario in science speaks for the reductionist under this interpretation. Though it's impossible to calculate the absolute cases of non-reductive explanations occurring in science, a cursive look at scientific

practice can tell that a large portion of scientific explanations proceeds in a non-reductive fashion, as suggested by multilevel modeling (Batterman 2013; Green 2013; O' Malley et al. 2014; Green and Batterman 2017; Mitchell and Gronenborn 2017). Moreover, even in areas such as physics which was regarded as a paradigm for the reductionist stance, progressive explanatory reduction doesn't always happen (Green and Batterman 2017; Batterman 2017).

In sum, we have shown that the in-principle argument fails for it neither offers help in understanding how science proceeds if it's construed as implying a logical possibility, nor goes in tune with scientific practice if it's construed as implying real possibilities.

5.2. Top-down causation

In Section 3 we have shown that there is a cross-level interaction taking the form that higher-level features may impact lower-level features. A worry arises: Does this imply top-down causation?

My answer to this question is twofold. First, it's clear that this short essay isn't aimed to engage in the philosophical debate about whether, and in what sense, there exists top-down causation (see Craver and Bechtel 2007; Kaiser 2015; Bechtel 2017). Second, what we can do now is to show that the cross-level interaction is a clear and well-defined concept in multilevel modeling. It unambiguously means the constraints on the lower-level processes exerted by the higher-level parameters (Green and Batterman 2017). In our multilevel modeling

discussed in Section 3, we have shown that group-level features may impact some individual-level features through the way that each group possesses its own feature relevant to explaining the differences at the individual-level across groups. This idea is incorporated into the mixed-effects model by assigning some explanatory variables to the group-level and a cross-level interaction term to the model.

The idea of cross-level-interaction-as-constraint is widely accepted in multilevel modeling broadly construed, where constraint is usually expressed in the form of initial and/or boundary conditions. For example, in modeling cardiac rhythms, due to "the influences of initial and boundary conditions on the solutions of the differential equations used to represent the lower level process" (Noble 2012, 55; Cf. Green and Batterman 2017, 32), a model cannot simply narrowly focus on the level of proteins and DNA but must also consider the levels of cell and tissue working as constraints. The same story happens in cancer research, where scientists are advocating the idea that tumor development can be better understood if we consider the varying constraints exerted by tissue (Nelson and Bissel 2006; Shawky and Davidson 2015; Cf. Green and Batterman 2017, 32).

6. conclusion

This essay has shown that no-reductive explanations involving many levels predominate in areas where the systems under consideration exhibit a hierarchical structure. These explanations violate the fundamental pillars of explanatory

reductionism: single level preference and lower-level obsession. Traditional single-level reductive approaches fall short of capturing systems of this kind because they face the challenges of committing either the atomistic or ecological fallacy.

References

Batterman, Robert. 2013. The "Tyranny of Scales." In *The Oxford Handbook of Philosophy of Physics*, ed. Robert Batterman, 255-286. Oxford: Oxford University Press.

———. 2017. "Autonomy of Theories: An Explanatory Problem." *Noûs* 1-16.

Bechtel, William. 2010. "The Downs and Ups of Mechanistic Research: Circadian Rhythm Research as an Exemplar." *Erkenntnis* 73:313–328.

———. 2017. "Explicating Top-Down Causation Using Networks and Dynamics." *Philosophy of Science* 84:253–274.

Bickle, John. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Brigandt, Ingo. 2010. "Beyond Reductionism and Pluralism: Toward an Epistemology of Explanatory Integration in Biology." *Erkenntnis* 73 (3): 295-311.

———. 2013a. "Explanation in Biology: Reduction, Pluralism, and Explanatory Aims." *Science and Education* 22:69–91.

———. 2013b. "Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:461–465.

Craver, Carl. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, Carl, and William Bechtel. 2007. "Top-down Causation without Top-Down Causes." *Biology and Philosophy* 22:547–563.

Freedman, David. 1999. "Ecological Inference and the Ecological Fallacy." In *International Encyclopedia of the Social and Behavioral* Sciences, vol. 6, ed. Neil Smelser, and Paul Baltes, 4027–4030. New York: Elsevier.

Green, Sara. 2013. "When One Model Isn't Enough: Combining Epistemic Tools in Systems Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:170–180.

Green, Sara, and Robert Batterman. 2017. "Biology Meets Physics: Reductionism and Multi-Scale Modeling of Morphogenesis." *Studies in History and Philosophy of Biological and Biomedical Sciences* 61:20–34.

Gross, Fridolin, and Sara Green. 2017. "The Sum of the Parts: Large-Scale Modeling in Systems Biology." *Philosophy, Theory, and Practice in Biology* 9: (10).

Heck, Ronald, and Scott Thomas. 2015. *An Introduction to Multilevel Modeling Techniques* (3rd Edition). New York: Routledge.

Hull, David. 1972. "Reductionism in Genetics—Biology or Philosophy?" *Philosophy of Science* 39 (4): 491-499.

Hüttemann, Andreas, and Alan Love. 2011. "Aspects of Reductive Explanation in Biological Science: Intrinsicality, Fundamentality, and Temporality." *British Journal for the Philosophy of Science* 62 (3): 519-549.

Kaiser, Marie. 2015. *Reductive Explanation in the Biological Sciences*. Springer.

Knuuttila, Tarja. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Studies in History and Philosophy of Science Part A* 42:262–271.

Luke, Douglas. 2004. *Multilevel Modeling*. London: SAGE Publications, Inc.

Maxwell, Sophie, Katherine Reynolds, Eunro Lee, et al. 2017. “The Impact of School Climate and School Identification on Academic Achievement: Multilevel Modeling with Student and Teacher Data.” *Frontiers in Psychology* 8:2069.

Mitchell, Sandra. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

———. 2009. *Unsimple Truths: Science, Complexity, and Policy*. Chicago: University of Chicago Press.

Nezlek, John. 2008. “An Introduction to Multilevel Modeling for Social and Personality Psychology.” *Social and Personality Psychology Compass* 2/2 (2008):842–860.

Noble, Daniel. 2012. “A Theory of Biological Relativity: No Privileged Level of Causation.” *Interface Focus* 2(1):55–64.

O’Malley Malley, Ingo Brigandt, Alan Love, et al. 2014. “Multilevel Research Strategies and Biological Systems.” *Philosophy of Science* 81:811–828.

Rosenberg, Alex. 2006. *Darwinian Reductionism, or How to Stop Worrying and Love Molecular Biology*. Chicago: University of Chicago Press.

Sarkar, Sahotra. 1998. *Genetics and Reductionism*. Cambridge: Cambridge University Press.

Schagen, I. P. 1990. “Analysis of the Effects of School Variables Using Multilevel Models.” *Educational Studies* 16:61–73.

Shawky, Joseph, and Lance Davidson. 2015. “Tissue Mechanics and Adhesion during Embryo Development.” *Developmental Biology* 401(1):152–164.

Sober, Elliot. 1999. “The Multiple Realizability Argument against Reductionism.” *Philosophy of science* 66:542–564.

Wang, Yau-De, and Hui-Hsien Hsieh. 2012. “Toward a Better Understanding of the Link Between Ethical Climate and Job Satisfaction: A Multilevel Analysis.” *Journal of Business Ethics* 105:535–545.

Waters, C. Kenneth. 2008. “Beyond Theoretical Reduction and Layer-Cake Antireduction: How DNA Retooled Genetics and Transformed Biological Practice”. In *The Oxford Handbook of Philosophy of Biology*, ed. Michael Ruse, 238-262. New York: Oxford University Press.

Weber, Marcel. 2005. *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

The Universe Never Had a Chance

C. D. McCoy[*†]

1 March 2018

Abstract

Demarest asserts that we have good evidence for the existence and nature of an initial chance event for the universe. I claim that we have no such evidence and no knowledge of its supposed nature. Against relevant comparison classes her initial chance account is no better, and in some ways worse, than its alternatives.
Word Count: 4712

1 Introduction

Although cosmology, the study of the universe's evolution, has largely become a province of physics, philosophical speculation concerning cosmogony, the study of the origin of the universe, continues up to the present. Certainly, many believe that science has settled this too by way of the well-known and well-confirmed big bang model of the universe. According to the big bang account the universe began in a extremely hot, dense state, composed of all the different manifestations of energy that we know. Indeed, time itself began with the big bang. Yet, properly speaking, the universe's past singularity is not some event in spacetime according to the general theory of relativity. In cosmological models this hot dense state called the big bang is generally understood instead as just a very early stage of the universe's evolution, i.e. properly a part of cosmology and not cosmogony. While we may be highly confident that the entire big bang story is correct back to a very early time, our confidence should at some point decrease as we near the supposed "first moment". Thus there remains world enough and time to engage in traditional philosophical and scientific speculations about cosmogony and cosmology alike. Were there previous stages to the universe? What brought the universe into existence? What was the character of this initial happening (should it in fact exist)?

The ubiquity of probabilities in modern physical theories, e.g. quantum mechanics and statistical mechanics, has led some to wonder as well how chance should fit into our

[*]**Acknowledgements:** Pending.

[†]School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK. email: casey.mccoy@ed.ac.uk

cosmogonical worldview. In this vein, Demarest (2016) argues that the probabilities of all events in a(n ostensibly) deterministic universe can be derived from an initial chance event and, what's more, that "we have good evidence of its existence and nature." In this paper I aim to dispute these latter claims. I argue that we do not have any evidence at all of an initial chance event in a big bang universe as described above, much less of its nature. What we rather have in Demarest's account is just a particular way of interpreting probabilistic theories, where all probabilities are taken to derive from ontic chances pertaining to the particular genesis of the relevant physical system, e.g. the universe as a whole. I claim that this interpretation, while coherent, should be disfavored in cosmology—we should rather say that *the universe never had a chance*.[1] Along the way I will make several clarifying remarks concerning the relation of chance and determinism, cosmological probabilities, and alternative interpretations of statistical and quantum mechanics.

2 Chance and Determinism in Physical Theory

By the *world* metaphysicians usually mean something like "the maximally inclusive entity whose parts are all the things that exist." Of course terminology varies. This particular rendering comes from Schaffer (2010, 33), who instead chooses to call this entity the *cosmos*. Cosmologists do not usually call their object of study the cosmos; more commonly they say that they study the *universe*. In *Cosmology: The Science of the Universe*, Harrison explicitly notes the philosophical and historical dimensions of the world taken in its broadest sense, designating this world as a whole the *Universe*. Cosmology, according to Harrison, is the study of universes, by which he means particular models of the Universe (Harrison, 2000, Ch. 1). Cosmological models are the particular concern of physical cosmologists; they are physical models of the Universe, which describe especially its large-scale structure and the evolution thereof.

In what follows I employ these terminologies in the following way. By the *world* I designate the locus of (principally) metaphysical questions concerning the Universe. Is the world deterministic? Is it chancy? By the *universe* I designate the locus of principally physical questions concerning the Universe. How did the big bang universe begin? How will it end? These are questions to which the big bang model should provide an answer.

I do not mean, of course, to introduce an admittedly arbitrary distinction between science and metaphysics by differentiating universes and worlds. Indeed, when one asks whether the world is deterministic, many metaphysicians of science would look first to models of the Universe to help decide the question. Wüthrich for example remarks, matter-of-factly, that "this metaphysical question deflates into the question of whether our best physical *theories* entail that the world is deterministic or indeterministic" (Wüthrich, 2011, 366).

[1] There are several senses, in fact, in which this claim is true. Cosmology suggests that the inevitable fate of the universe is to become ever more sparse and empty through the accelerated expansion of space under the influence of dark energy.

Indeed, many discussions of determinism adopt the approach mentioned by Wüthrich. Let *determinism* denote the thesis that the world is deterministic. Then, following for example (Lewis, 1983, 360), a world is *deterministic* if and only if the laws of that world are deterministic. To determine whether the laws of the universe are deterministic, we must look to our theories of which those laws are part and ask whether those laws taken together should be considered deterministic. It is by no means a straightforward matter to decide whether a given physical theory is deterministic of course. Even the classic example of deterministic physics, Newtonian mechanics, admits many counterexamples against its putative determinism (Earman, 1986; Norton, 2008). General relativity as well seemingly permits indeterministic phenomena in the form of causal pathologies (closed timelike curves) (Earman, 1995) and, if the hole argument is to be believed, is hopelessly rife with indeterminism (Earman and Norton, 1987).

Although classical theories like classical mechanics and general relativity are nevertheless debatably deterministic, surely probabilistic theories like quantum mechanics are properly characterized as indeterministic (at least so long as the probabilities involved are objective features of the world). Yet various interpretations of probabilistic theories seek to avoid indeterminism even here, where it seems unassailable, by characterizing probabilities as merely epistemic or subjective, or else by presenting them as fully deterministic theories (as in the Bohmian interpretation of quantum mechanics). Philosophers have raised serious concerns, however, over how one can truly understand probabilities in deterministic theories, an issue that has been termed the "paradox of deterministic probabilities" (Loewer, 2001; Winsberg, 2008; Lyon, 2011) in statistical mechanics, since objective probabilities seem to entail indeterminism necessarily.

The most well-known and successful reconciliation of chance and determinism in the context of statistical mechanics is defended by Loewer (2001). It is seldom recognized by interpreters, however, that there is no reconciliation in the sense of simultaneous compatibility between chance and determinism. The world cannot both be chancy and deterministic as a matter of metaphysical fact. As Lewis writes, "to the question of how chance can be reconciled with determinism, or to the question of how disparate chances can be reconciled with one another, my answer is: *it can't be done* (Lewis, 1986, 118). This is because chance entails indeterminism, the contrary of determinism. Thus, insofar as the probabilities of statistical mechanics and quantum mechanics are objective, these theories are indeterministic theories. Loewer's account actually shows us how deterministic laws can co-exist with indeterministic laws within a theory. The source of all probabilities in statistical mechanics, according to Loewer, is in an initial chance distribution over microscopic states of affairs. After the initial time these states of affairs evolve deterministically. Note that although for almost all times evolution is deterministic, it is not so at all times. There is an initial chance event, which is where the indeterminism of the theory appears. A deterministic theory is, recall, a theory whose laws are deterministic, not a theory whose laws are mostly deterministic or operate deterministically for almost all times.

Loewer's account is also presented in terms of Humean chances, so he does not believe

these chances and laws actually exist. According to the modern Humean, they merely are the result of the best systematizations of the occurrent facts, in keeping with Lewis's "best systems account" of laws and chances. Demarest, however, offers a small tweak to Loewer's Humean account by invoking a "robustly metaphysical account of chance" (Demarest, 2016, 256). She claims that such chances are compatible with determinism, and indeed they are when, as said, compatibility is understood to pertain to the co-existence of indeterministic and deterministic laws in a single theory—which, however, do not operate at the same time.[2]

Demarest's central claims are that this initial chance event exists and that we have good evidence for it. I dispute these claims in the remainder of the paper.

To begin, it is not so clear what exactly Demarest takes the evidence for the initial chance event to be. She does contrast the evidential position of her view with the Humean view of Loewer, claiming that, "for the Humean, the statistical patterns in the world are not evidence of an initial chance event" (Demarest, 2016, 261)—presumably this is so because Humeans reject the metaphysics of chance for the usual Humean reasons. One might suppose, then, that she believes that statistical patterns in the world are evidence of an initial chance event for all those who do not share the Humeans ontological worries. Let us accept, for the moment then, that statistical patterns may be *some* evidence for the existence of chances, for it is difficult to see what other evidence there might be for an initial chance event. In that case, on what grounds might we say that statistical patterns are good evidence for initial chances? I consider a series of three salient contrast classes.

First, do statistical patterns in data provide good evidence for indeterministic (i.e. chancy) theories *rather than deterministic theories*? It would seem that the answer is: not necessarily. (Werndl, 2009), for example, argues for the observational equivalence of indeterministic theories and deterministic theories. If one could contrive a fully deterministic theory that reproduces the same statistical patterns of the relevant phenomena observed in nature, then it would seem that such patterns provide no better evidence for the indeterministic theory than the deterministic one. However, since the theories under discussion, statistical mechanics and quantum mechanics, are generally characterized as indeterministic, let us flag but set aside the possibility of fully deterministic alternatives to them.

So, second, do statistical patterns provide good evidence for initial chances *rather than non-initial chances*? It would seem that the answer is firmly: no. There is a variety of ways one could implement chances into a probabilistic theory like statistical mechanics. All one must do, as Loewer shows us by example, is neatly separate when the indeterministic laws are operative and when the deterministic laws are operative. Loewer chooses to locate all the indeterminism in one place—the initial time—but one could equally locate it at another time, at many times, or even all times. Statistical mechanics does not wear its interpretation on its sleeve, just as quantum mechanics does not decide between solutions of the measurement problem, whether initial chances as in Bohmian mechanics or collapse

[2]Still, it is worth emphasizing that her claim that her account applies to deterministic worlds is false, for chancy worlds are not deterministic.

dynamics as in GRW (discrete time collapses) or CSL (continuous collapses). Unless there are evidential reasons to favor one implementation of indeteterministic probabilities over the others, there is not good evidence for an initial chance event. Certainly statistical patterns in nature will not do so.

Third, do statistical patterns provide good evidence for "robustly metaphysics" chances *rather than Humean chances*? It seems as if this might Demarest's intended contrast class, since much of the discussion in the paper concerns the Humean account. I will have something to say about the relative merits of Demarest's non-Humean account and Loewer's Humean account at the end of the next section. In any case though, it does not seem as if statistical patterns decide the matter in Demarest's mind, for she repeatedly demurs in the face of Humean responses to the considerations she raises, claiming only to offer an alternative "for philosophers who are antecedently sympathetic to governing laws of nature or powerful properties" (Demarest, 2016, 261-2). She finds it "plausible to think of the universe as having an initial state and as producing subsequent states in accordance with the laws of nature (some of which may be chancy)" (Demarest, 2016, 261). Such metaphysical intuitions are not grounded on observations of statistical patterns. Statistical patterns do not have any evidential bearing on the metaphysical dispute between the Humean and non-Humean.

Therefore, based on my canvassing of relevant alternatives, I conclude that we in fact do not have good evidence for an initial chance event, where evidence is interpreted in terms of statistical patterns (or in any usual sense of the term "evidence"). At best we have a motivation to attend to indeterministic theories when our evidence displays statistical patterns. It is another matter entirely to decide how to implement probabilities in that theory.

That said, Demarest's reasoning could be interpreted at points as invoking explanatory considerations as justification for the initial chance interpretation. Insofar as one considers "what justifies" as constituting evidence, perhaps these explanatory considerations should be counted as evidence.[3] Nevertheless, it does not look, on the face of it, like we have good evidence for an initial chance event still. Repeating the three cases considered before: deterministic and chancy theories can both serviceably explain statistical evidence; alternative implementations of chance in interpretations of indeterministic theories explain statistical evidence equally well; Humean and non-Humean metaphysics each render a story for how statistical patterns come about (merely subjective intuitions notwithstanding). Without explicit explanatory reasons to prefer one of these alternatives to the other, reasons lacking in Demarest's argument, good evidence (in this wider sense) for an initial chance event remains elusive.

[3]There are obvious dangers with going to far in this direction. Suppose that the Supreme Being explains all. Then it would appear that we have very good evidence of Its existence, which is obviously absurd.

3 Chance and Determinism in Systems of the World

In the previous section I gave reasons to doubt Demarest's claims about an initial chance event and our evidence for it. I disputed especially that we have evidence for it and did so by comparing it to alternatives of three different kinds. In the first case I characterized the issue (in part) as a matter of theory choice, namely of choosing between an indeterministic and deterministic theory. In the second case I characterized the issue as a matter of theory interpretation, namely of interpreting between different ways of implementing probability in a theory that does not decide one way or another on how this must be done. In the third case I characterized the issue as a matter of metaphysics, namely of deciding between the ontological status of chances.

In this section I consider more broadly whether there are any reasons to favor Demarest's interpretation, in particular in the sense of the just given second characterization of the issue. The question is whether the world should be thought to have an initial chance event, when one might consider that it is chancy in various other ways, e.g. its laws of evolution themselves are always probabilistically indeterministic.

First of all, it is worth mentioning that from the point of view given by the contemporary standard model of cosmology this question is moot. The so-called ΛCDM model, a development of the older standard big bang model, is a model of the general theory of relativity, a theory which makes use of no probabilities at all in its basic description of gravitating systems (including the universe). In this different sense it is also true that the universe never had a chance.

Demarest is not particularly interested in cosmology or the universes of general relativity however. She is concerned with probabilistic theories like classical statistical mechanics and quantum mechanics as applied to the world at large. We should, that is, imagine a statistical mechanical universe or a quantum mechanical universe (never minding that no concrete such model exists in physics that describes our universe) as a conceptual possibility when asking metaphysical questions about the world. Given the different ways of implementing probabilities in such a universe, we should ask whether one way is preferable to the others.

I should point out that this is not Demarest's question, for she explicitly restricts attention to "deterministically evolving worlds". Of course these worlds are not actually deterministic so long as the probabilities involved are chances. Nevertheless, unaffected by that fact is one of her central points: "that positing just one initial chance event can justify the usefulness and explain the ubiquity of nontrivial probabilities to epistemic agents like us, even if there are no longer any chance events in our world"(Demarest, 2016, 249). I say: so can a lot of other ways of conceiving chance in these theories. It is therefore necessary to compare them if we are to take Demarest's (and Loewer's) account seriously.

For present purposes, I am happy to agree with Demarest that the initial chance account can indeed justify and explain nontrivial probabilities used to describe subsystems of the universe.[4] But is it a good explanation? Is it worth believing?

[4]Notwithstanding pressure to move in this "global" direction in statistical mechanics (Callender, 2011)

The initial chance account invites the oft-invoked (in cosmology) picture of the (blind and unskilled) Creator throwing a dart (Wald, 2006, 396) or pointing a pin (Penrose, 1989, 442) at the set of possible universes, thereby picking out the initial conditions of the universe. That such pictures are intended as pejorative jabs at dubious metaphysics is plain. A mere picture is hardly an objection, of course, so what is it that seems problematic about initial chances for the universe? Could it not be the best cosmogonical story of our universe, that is, that a matter of chance determined its actualization out of a vast range of possibilities that could have been actualized had only their sisal been struck?

Intuition suggests that this just is not a serious, satisfying story for how the world could be. The probabilities of events in the actual world would derive ultimately from the probabilities for the actualization of our world. But why should we not just assume that the world started in the state that it did, with probability one or with certainty? Presumably the response of the initial chance advocate is that in that case we would lose the justification and explanation of subsystem probabilities. Yet is there anything to lose, if this metaphysical explanation is epistemically untrustworthy? How can we come to know these ultimate probabilities of other worlds? Is the metaphysical story sufficiently complete even? How could the probabilities of other worlds matter for what happens in *our* world?

I am willing to grant that these questions do have some answer, for what strikes me as a more serious difficulty is the following. Insofar as they are objective and justified, the probabilities agents like us use for specific events in subsystems of the world must be epistemic probabilities. On Demarest's (and Loewer's) account all such epistemic probabilities derive from initial epistemic probabilities for different initial conditions of the world. How is it that these probabilities obtain their needed objectivity and justification, and hence explanatory power? According to Demarest it is because they accord with the actual chances. However, what has one achieved by invoking "actual chances" at this stage? Although these chances do not merely have a *virtus dormitiva* per se, "just so" stories like this surely make the explanatory credentials of chances suspect. Does one dare invoke a transcendental argument or thump the realist table to defend their objectivity?

If we were somehow forced to adopt the initial chance explanation of epistemic probabilities, then we might swallow whatever dubious metaphysics attendant to it. If there were reasonable alternatives, however, should we not prefer them? And indeed there are other interpretive options available. Locating the chances at another time (or even "outside the universe") constitutes one set of possibilities, but they obviously suffer from the same awkwardness as the initial chance account. Another is based on the idea that chancy behavior occurs at discrete time intervals. One finds this idea in the orthodox Copenhagen and other collapse interpretations of quantum mechanics for example. One might be uneasy with the invocation of chancy behavior at potentially ill-defined times in such interpretations, and even with their postulation of two dynamical laws of nature, a deterministic one and an indeterministic one (although it is a feature of the initial chance account as well). However one at least avoids a commitment to chance figuring into

(and quantum mechanics) in order to justify and explain probabilities in subsystems of the universe, serious reservations about whether doing so is itself justified are advanced by, inter alia, Earman (2006).

cosmogenesis and also the questionable leap to objectivity in agential probabilities, since chances in these interpretations are physical processes that happen within the universe, whether as part of the general evolution of the universe or tied to the evolution of individual systems.

Another possibility is suggested by continuing this line of thought, i.e. of spreading chanciness out further in time. Instead of chancy behavior at discrete intervals, why not suppose that it occurs continuously? In quantum mechanics this idea is implemented in some interpretations, such as continuous spontaneous localization, and in statistical mechanics there are various stochastic dynamics approaches. Advantages of this idea are that one has a single law of evolution, an indeterministic one, and, again, one does not make chanciness a matter of cosmogenesis. What disadvantage? To some that it makes the world rife with indeterminism. Yet who is afraid of indeterminism? It surely does not mean anything goes, nor does it threaten the possibility of knowledge of the world (although there are limits to what we can know). Besides, by accepting quantum mechanics (or even statistical mechanics) we have already let indeterminism in the door in physics.

When we look at the interpretations available for a world governed by probabilistic laws, in every case the alternatives to the initial chances view therefore appear preferable. Indeed, it would seem that only one who demands that the world be as deterministic as possible could favor the initial chances view, but it is hard to see what motivation there could be for that demand. I therefore conclude, in a final sense, that *the universe never had a chance*.

That said, I emphasize that this judgment applies only to the case where we treat the universe as a statistical mechanical system or quantum mechanical system. In other words, the world is the universe, our world-metaphysics is our universe-metaphysics. The considerations leading to this conclusion change shape somewhat when we confine the application of our theories to systems describable by those theories. The initial chance account is far less dubious when attached to individual statistical mechanical systems and not automatically to the universe at large. Indeed, it could well be that the initial conditions of similar systems are best treated as randomly distributed, for here we do have empirical evidence that this interpretation can be used to explain—unlike with the universe, where we have but one system.

There is, as noted, sometimes pressure to globalize our theories, especially in the case of statistical mechanics. If we ask what accounts for the randomness in initial conditions of a particular class of systems, it is natural to look at larger systems that contain them. If we find that these systems have random initial conditions, then we continue to expand our scope, ultimately reaching the "maximally inclusive entity whose parts are all the things that exist." This globalization of statistical mechanics is the kernel of the so-called imperialism of (Albert, 2000) and Loewer. If we are right to feel this pressure to interpret the world at large in the same terms as individual physical systems, then there is concomitant pressure to hold the same interpretive of chance in both cases. I have argued, however, that the intuitive considerations vary somewhat, at least with respect to the initial chance account. Is this reason to disfavor it in the case of individual systems? Or is our confidence in its applicability for individual systems sufficient to overcome any hesitation at

accepting it for the universe? My inclination is to answer "yes" and "no", but I offer no grounds for the preference here. I do believe that metaphysicians of science should care about considerations like this, however, having to do with the relation of subsystem and universe, for often enough what seems right in one context is questionable in the other.

I close this section with a brief comment on the relation of Loewer's and Demarest's accounts. As I argued above, empirical evidence and explanatory considerations do not favor one over the other, since they account for empirical evidence in essentially the same way. The central difference is whether chances are understood as reducible to other facts, hence not part of the fundamental ontology of the world, or as "robustly metaphysical", in which case they are. The problems Demarest mentions for the Humean view—past events may have nontrivial chances, the chance of an event depends on what one knows, worlds with identical frequencies cannot have different chances, etc.—are surely not problems when viewed properly through the Humean lens. However, whereas the problem I raise for the initial chance view, concerning the explanatory credentials and justification for the posit of initial chances, threatens Demarest's account, it will not worry the Humean of Loewer's stripe, for these initial chances do not exist for the Humean. Humean chances do not produce or generate any actual states of affairs. Of course one may raise the usual complaint against the Humean, that there is a circularity in the Humean account involving descriptions explaining themselves, and others besides. I do not care to enter into this debate here of course. I only wish to point out that my argument about how chance can fit into a cosmogonical worldview appears to give some reason to favor the Humean account in this particular context.

4 Conclusion

In this paper I considered whether we should think that the world had one chance, as claimed by Demarest. First I considered her claim that we have good evidence that an initial chance event occurred by contrasting it with relevant classes of alternatives. I argued that evidence neither favors a chancy theory over a chanceless theory, nor initial chances over other implementations of chances, nor metaphysically robust chances over Humean chances. I concluded, therefore, that we do not have good evidence to adopt the initial chance account.

I then considered whether there were other reasons to favor or disfavor the initial chance account. I argued that the dubious nature of worldly chances provides a strong impulse to look for other accounts that do not make chance a matter of cosmogenesis. The other implementations did not suffer from this defect, so I suggested that from a cosmogonical perspective they should be preferred. But the relation of the universe and its subsystems makes a demand to have a consistent interpretation. As the initial chance account looks favorable on the subsystem level (to many) and not on the universe's level (as I argued), there remains a significant metaphysical tension to be resolved.

References

Albert, D. (2000). *Time and Chance*. Cambridge, MA: Cambridge, MA: Harvard University Press.

Callender, C. (2011). The past histories of molecules. In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 83–113. Oxford: Oxford University Press.

Demarest, H. (2016). The universe had one chance. *Philosophy of Science 83*(2), 248–264.

Earman, J. (1986). *A Primer on Determinism*. Dordrecht: D. Reidel Publishing Company.

Earman, J. (1995). *Bangs, Crunches, Whimpers, and Shrieks*. Oxford: Oxford University Press.

Earman, J. (2006). The "past hypothesis": Not even false. *Studies in History and Philosophy of Modern Physics 37*, 399–430.

Earman, J. and J. Norton (1987). What price spacetime substantivalism? the hole story. *British Journal for the Philosophy of Science 38*, 515–525.

Harrison, E. (2000). *Cosmology: the science of the universe* (2nd ed.). Cambridge: Cambridge University Press.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Lewis, D. (1986). *Philosophical Papers*, Volume 2. Oxford: Oxford University Press.

Loewer, B. (2001). Determinism and chance. *Studies in History and Philosophy of Modern Physics 32*, 609–620.

Lyon, A. (2011). Deterministic probability: neither chance nor credence. *Synthese 182*, 413–432.

Norton, J. (2008). The dome: An unexpectedly simple failure of determinism. *Philosophy of Science 75*, 786–798.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

Schaffer, J. (2010). Monism: The priority of the whole. *The Philosophical Review 119*, 31–76.

Wald, R. (2006). The arrow of time and the initial conditions of the universe. *Studies in History and Philosophy of Modern Physics 37*, 394–398.

Werndl, C. (2009). Are deterministic descriptions and indeterministic descriptions observationally equivalent? *Studies in History and Philosophy of Modern Physics 40*, 232–242.

Winsberg, E. (2008). Laws and chances in statistical mechanics. *Studies in History and Philosophy of Modern Physics 39*, 872–888.

Wüthrich, C. (2011). Can the world be shown to be indeterministic after all? In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 365–389. Oxford: Oxford University Press.

Draft paper for the symposium *Mechanism Meets Big Data: Different Strategies for Machine Learning in Cancer Research* to be held at *PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association* (Seattle, WA; 1-4 Nov 2018).

MECHANISTIC MODELS AND THE EXPLANATORY LIMITS OF MACHINE LEARNING

Emanuele Ratti[1], University of Notre Dame

Ezequiel López-Rubio, Universidad Nacional de Educación a Distancia, University of Málaga

Abstract

We argue that mechanistic models elaborated by machine learning cannot be explanatory by discussing the relation between mechanistic models, explanation and the notion of intelligibility of models. We show that the ability of biologists to understand the model that they work with (i.e. intelligibility) severely constrains their capacity of turning the model into an explanatory model. The more a mechanistic model is complex (i.e. it includes an increasing number of components), the less explanatory it will be. Since machine learning increases its performances when more components are added, then it generates models which are not intelligible, and hence not explanatory.

1. INTRODUCTION

Due to its data-intensive turn, molecular biology is increasingly making use of machine learning (ML) methodologies. ML is the study of generalizable extraction of patterns from data sets starting from a problem. A problem here is defined as a given set of input variables, a set of outputs which have to be calculated, and a sample (previously input-output pairs already observed). ML calculates a quantitative relation between inputs and outputs in terms of a predictive model by learning from an already structured set of input-output pairs. ML is expected to increase its performances when the complexity of data sets increase, where complexity refers to the number of input variables and the number of samples. Due to this capacity to handle complexity, practitioners think that ML is potentially able to deal with biological systems at the macromolecular level, which are notoriously complex. The development of ML has been proven useful not just for the

[1] mnl.ratti@gmail.com

complexity of biological systems *per se*, but also because biologists now are able to generate an astonishingly amount of data. However, we claim that the ability of ML to deal with complex systems and big data comes at a price; *the more ML can model complex data sets, the less biologists will be able to explain phenomena in a mechanistic sense*.

The structure of the paper is as follows. In Section 2, we discuss mechanistic models in biology, and we emphasize a surprising connection between explanation and model complexity. By adapting de Regt's notion of pragmatic understanding (2017) in the present context, we claim that if a how-possibly mechanistic model can become explanatory, then it must be intelligible to the modeler (Section 2.2, 2.3 and 2.4). Intelligibility is the ability to perform precise and successful material manipulations on the basis of the information provided by the model about its components. The results of these manipulations are fundamental to recompose the causal structure of a mechanism out of a list of causally relevant entities. Like a recipe, the model must provide instructions to 'build' the phenomenon, and causal organization is fundamental in this respect. If a model is opaque to these organizational aspects, then no mechanistic explanations can be elaborated. By drawing on studies in cognitive psychology, we show that the more the number of components in a model increases (the more the model is complex), the less the model is intelligible, and hence the less an explanation can be elaborated.

Next, we briefly introduce ML (Section 3). As an example of ML application to biology, we analyze an algorithm called PARADIGM (Vaske et al 2010), which is used in biomedicine to predict clinical outcomes from molecular data (Section 3.1). This algorithm predicts the activities of genetic pathways from multiple genome-scale measurements on a single patient by integrating information on pathways from different databases. By discussing the technical aspects of this algorithm, we will show how the algorithm generates models which are more accurate as the number of variables included in the model increases. By variables, here we mean biological entities included in the model and the interactions between them, since those entities are modeled by variables in PARADIGM.

In Section 4 we will put together the results of Section 2 and 3. While performing complex localizations more accurately, we argue that an algorithm like PARADIGM makes mechanistic models so complex (in terms of the number of model components) that no explanation can be constructed. In other words, ML applied to molecular biology undermines biologists' explanatory abilities.

2. COMPLEXITY AND EXPLANATIONS IN BIOLOGY

The use of machine learning has important consequences for the explanatory dimension of molecular biology. Algorithms like PARADIGM, while providing increasingly accurate localizations, challenge the explanatory abilities of molecular biologists, especially if we assume the account of explanation of the so-called mechanistic philosophy (Craver and Darden 2013; Craver 2007; Glennan 2017). In order to see how, we need to introduce the notion of mechanistic explanation, and its connection with the notion of intelligibility (de Regt 2017).

2.1 Mechanistic explanations

Molecular biology's aim is to explain how phenomena are produced and/or maintained by the organization instantiated by macromolecules. Such explanations take the form of mechanistic descriptions of these dynamics. As Glennan (2017) succinctly emphasizes, mechanistic models (often in the form of diagrams complemented by linguistic descriptions) are vehicles for mechanistic explanations. Such explanations show how a phenomenon is produced/maintained and constituted by a mechanism – mechanistic models explain by explaining *how*. As Glennan and others have noticed, a mechanistic description of a phenomenon looks like what in historical narrative is called *causal narrative*, in the sense that it "describes sequences of events (which will typically be entities acting and interacting), and shows how their arrangement in space and time brought about some outcome" (Glennan 2017, p 83). The main idea is that we take a set of entities and activities to be causally relevant to a phenomenon, and we explain the phenomenon by showing how a sequence of events involving the interactions of the selected entities produces and/or maintains the explanandum. In epistemic terms, it is a

matter of showing a chain of inferences that holds between the components of a model (e.g. biological entities). Consider for instance the phenomenon of restriction in certain bacteria and archaea (Figure 1). This phenomenon has been explained in terms of certain entities (e.g. restriction and modification enzymes) and activities (e.g. methylation). Anytime a bacteriophage invades one of these bacteria or archaea (from now on *host cells*), host cells stimulate the production of two types of enzymes, i.e. a restriction enzyme and a modification enzyme. The restriction enzyme is designed to recognize and cut specific DNA sequences. Such sequences, for reasons we will not expose here[2], are to be found in the invading phages and/or viruses. Hence, the restriction enzyme destroys the invading entities by cutting their DNA. However, the restriction enzyme is not able to distinguish between the invading DNA and the DNA of the host cell. Here the modification enzyme helps, by methylating the DNA of the host cell at specific sequences (the same that the restriction enzyme cuts), thereby preventing the restriction enzyme to destroy the DNA of the host cell. The explanation of the phenomenon of restriction is in terms of a narrative explaining how certain entities and processes contribute to the production of the phenomenon under investigation. The inferences take place by thinking about the characteristics of the entities involved, and how the whole functioning of the system can be recomposed from entities themselves.

[2] See for instance (Ratti 2018)

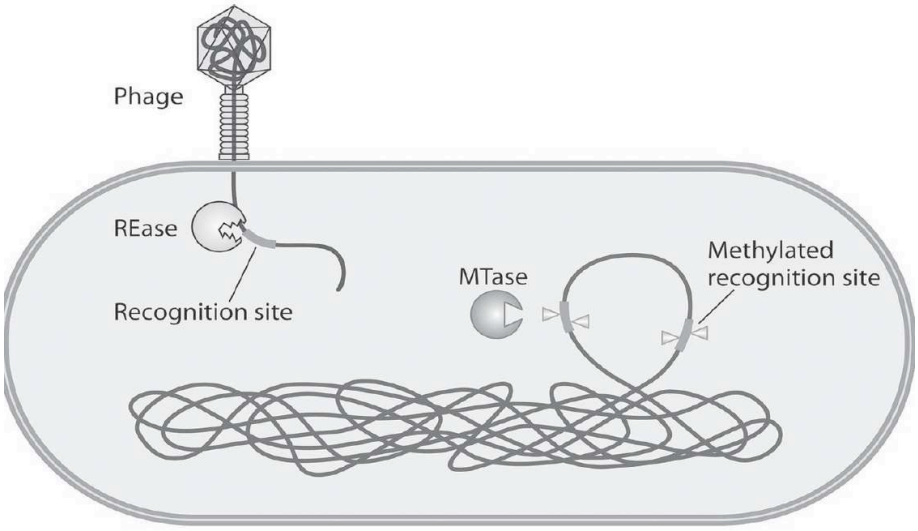


Figure 1. Mechanistic model of restriction. A phage enters a bacterium cell and sequences of its DNA are cleaved by a restriction enzyme (REase). Simultaneously, a modification enzyme (MTase) methylates a specific sequence in the DNA of host so that the restriction enzyme does not cleave the genome of the host too. Original figure taken from (Vasu and Nagaraja 2013).

2.2. Complexity of mechanistic models

Despite the voluminous literature on mechanistic explanation, there is a connection between models, *in fieri* explanations and the modeler that has not been properly characterized. In particular, mechanistic models should be intelligible to modelers in order to be turned into complete explanations. Craver noticed something like that when he states that his ideal of completeness of a mechanistic description (in terms of molecular details) should not be taken literary, but completeness always refer to the particular explanatory context one is considering. The reason why literary completeness is unattainable is because complete models will be of *no use* and completely *obscure* to modelers; "such descriptions would include so many potential factors that they would be *unwieldy for the purpose of prediction and control and utterly unilluminating to human beings*" (2006, p 360, emphasis added).

We rephrase Craver's intuitions by saying that *how-possibly models cannot be turned into adequate explanations if they are too complex*. We define complexity as *a function of the number of entities and activities (i.e. components of the model) that have*

to be coordinated in an organizational structure in the sense specified by mechanistic philosophers. This means that no agent can organize the entities and/or activities localized by highly complex models in a narration that rightly depicts the organizational structure of the *explanandum*. Therefore, very complex models which are very good in localization cannot be easily turned into explanations. Let us show why complex models cannot be turned into explanatory models in the mechanistic context.

2.3 Intelligibility of mechanistic models

The idea that agents cannot turn highly complex mechanistic models into explanations can be made more precise by appealing to the notion of *intelligibility* (de Regt 2017).

By following the framework of models as mediators (Morgan and Morrison 1999), de Regt argues that models are the way theories are applied to reality. Similar to Giere (2010), de Regt thinks that theories provide principles which are then articulated in the form of models to explain phenomena; "[t]he function of a model is to represent the target system in such a way that the theory can be applied to it" (2017, p 34). He assumes a broad meaning of explanation, in the sense that explanations are arguments, namely attempts to "answer the question of why a particular phenomenon occurs or a situation obtains (…) by presenting a systematic line of reasoning that connects it with other accepted items of knowledge" (2017, p 25). *Ça va sans dire*, arguments of the sort are not limited to linguistic items[3]. On this basis, de Regt's main thesis is that a *condition sine qua non* to elaborate an explanation is that the theory from which it is derived must be intelligible.

In de Regt's view, the intelligibility of a theory (*for scientists*) is "[t]he value that scientists attribute to the cluster of virtues (…) that facilitate the use of the theory for the construction of models" (p 593). This is because an important aspect of obtaining explanations is to derive models from theories, and to do that a scientist must use the theories. Therefore, if a theory possesses certain characteristics that make it easier to be used by a scientist, then the same scientist will be in principle more successful in deriving explanatory models. In (2015) de Regt extends this idea also to models in the sense that "understanding consists in being able to use and manipulate the model in order to make

[3] Mechanistic explanations are arguments, though not of a logical type

inferences about the system, to predict and control its behavior" (2015, p 3791). If for some reasons models and theories are not intelligible (to us), then we will not be able to develop an explanation, because we would not know how to use models or theories to elaborate one.

This idea of intelligibility of models and its tight connection with scientific explanation, can be straightforwardly extended to mechanistic models. Intelligibility of mechanistic models is defined by the way we *successfully* use them to explain phenomena. But how do we use models (mechanistic models in particular), and for what? Please keep in mind that whatever we do with mechanistic models, it is with explanatory aims in mind. Anything from predicting, manipulating, abstracting, etc is because we want an explanation. This is a view shared both by mechanistic philosophers but by de Regt as well, whose analysis of intelligibility is in explanatory terms.

First, highly abstract models can be used to build more specific models, as in the case of schema (Machamer et al 2000; Levy 2014). A schema is "a truncated abstract description of a mechanism that can be filled with descriptions of known component parts and activities" (Machamer et al 2000, p 16). For instance, consider the model of transcription. This model can be highly abstract where 'gene' stands for any gene, and 'transcription factor' stands for any transcription factor. However, we can instantiate such a schema in a particular experimental context by specifying which gene and which transcription factors are involved. The idea is that biologists, depending on the specific context they are operating, can instantiate experiments to find out which particular gene or transcription factor is involved in producing a phenomenon at a given time.

Next, mechanistic models can be used in the context of the *build-it test* (Craver and Darden 2013) with confirmatory goals in mind. Since mechanistic explanations may be understood as recipes for construction, and since recipes provide instructions to use a set of ingredients and instruments to produce something (e.g. a cake), then mechanistic models provide instructions to build a phenomenon or instructions to modify it in controlled ways because, after all, they tell us about the internal division of labor between entities causally relevant to producing or maintaining phenomena. This is in essence the build-it test as a confirmation tool; by modifying an experimental system on the basis of the 'instructions' provided by the model that allegedly explains such a phenomenon, we

get hints as to how the model is explanatory. If the hypothesized modifications produce in the 'real-world' the consequences we have predicted on the basis of the model, then the explanatory adequacy of the model is corroborated. The more the modifications suggested are precise, the more explanatory the model will be[4]. A first lesson we can draw is that *if a mechanistic model is explanatory, then it is also intelligible*, because it is included in the features of being explanatory mechanistically the fact that we can use the model to perform a build-it test.

The build-it test is also useful as a *tool to develop* explanations. Consider again the case of restriction in bacteria and a how-possibly model of this phenomenon based on a few observations. Let's say that we have noticed that when phages or viruses are unable to grow in specific bacteria, such bacteria also produce two types of enzymes. We know that the enzymes, the invading phages/viruses and restriction are correlated. The basic model will be as follows; anytime a phage or a virus invade a bacterium, these enzymes are produced, and hence the immune system of the bacterium must be related to these enzymes. We start then to instantiate experiments on the basis of this simple model. Such a model suggests that these enzymes must do something to the invading entities, but that somehow modify the host cell as well. Therefore, the build-it test would consist in a set of experiments to stimulate and/or inhibit these entities to develop our ideas about the nature of their causal relevance and their internal division of labor. *In fieri* mechanistic models suggest a range of instructions to 'build' or 'maintain' phenomena. These instructions are used to instantiate experiments to refine the model and make it explanatory. This is an example of what Bechtel and Richardson would call *complex localization* (2010, Chapter 6), and it is complex because the strategy used to explain the behavior of a system (immune system of host cells) is heavily constrained by empirical results of lower-levels. The how-possibly model affords a series of actions leading to a case of complex localization, when "constraints are imposed, whether empirical or theoretical, they can serve simultaneously to vindicate the initial localization and to develop it into a full-blooded mechanistic explanation" (Bechtel and Richardson 2010, p 125). Therefore, *if a how-possibly model can be turned into an explanatory model, then it*

[4] Please note that such a test, when involving adequate mechanistic explanations, is also the preferred way to teach students in text books, or also a way to provide instructions to reproduce the results of a peer-reviewed article

is intelligible, because the way we turn it into an explanatory model is by instantiating build-it tests.

A mechanistic model is therefore intelligible either when (a) it is a schema and we can instantiate such a model in specific contexts, or (b) when it affords a series of built-it test which are used either to corroborate its explanatory adequacy, or to make it explanatory. About (b), it should be noted that if we consider a mechanistic model as a narrative, then the model will be composed of a series of steps which influence each other in various ways. *Being able to use a model means being able to anticipate what would happen to other steps if I modify one step in particular*. This is not a yes/no thing. The model of restriction-modification systems is highly intelligible, because I know that if I prevent the production of modification enzymes I simultaneously realize that the restriction enzyme will destroy the DNA of the host cell. However, more detailed models will be less intelligible, because it would be difficult to simultaneously anticipate what would happen at each step by modifying a step in particular.

2.4 Recomposing mechanisms and intelligibility

In the mechanistic literature, the process of developing an explanatory model out of a catalogue of entities that are likely to be causally relevant to a phenomenon is called *recomposition of a mechanism* and it usually happens after a series of localization steps.

To recompose a mechanism, a modeler must be able to identify causally relevant entities and their internal division of labor. The idea is not just to 'divide up' a given phenomenon in tasks, but also a given task in subtasks interacting in the overall phenomenon, as it happens in complex localization (Bechtel and Richardson 2010). In the simplest case, researchers assume linear interactions between tasks, but there may be also non-linear or more complex type of interactions.

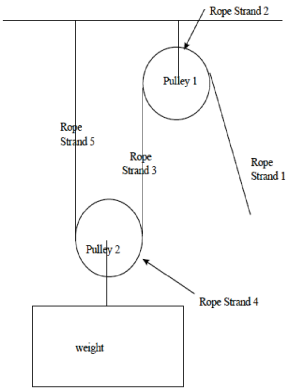
These reasoning strategies are usually implemented by thinking about these dynamics with the aid of *diagrams*. Diagrammatic representations usually involve boxes standing for entities (such as genes, proteins, etc) and arrows standing for processes of various sorts (phosphorylation, methylation, binding, releasing, etc). Therefore, biologists recompose mechanisms as mechanistic explanations by thinking about these diagrams,

and they instantiate experiments (i.e. built-it test) exactly on the basis of such diagrammatic reasoning.

Cognitive psychology and studies of scientific cognition have extensively investigated the processes of diagrammatic reasoning (Hegarty 2000; 2004; Nersessian 2008). Moreover, empirical studies have emphasized the role of diagrams in learning and reasoning in molecular biology (Kindfield 1998; Trujillo 2015). In these studies, diagrammatic reasoning is understood as a "task that involves inferring the behavior of a mechanical system from a visual-spatial representation" (Hegarty 2000, p 194). Hegarty refers to this process as *mental animation*, while Nersessian (2008) thinks about this as an instantiation of *mental modelling*. This is analogous to thinking about mechanistic models as narratives, namely being able to infer how a course of events, decomposed into steps, may change if we change one step in particular. Mental animation is a process of complex visual-spatial inference. Limits and capabilities of humans in such tasks depend on the cognitive architecture of human mind[5]. What Hegarty has found is that mental animation is *piecemeal*, in the sense that human mind does not animate the components of a diagram in parallel, but rather infer the motion of components *one by one*. This strategy has a straightforward consequence; in order to proceed with animating components, we should store intermediate results of inferences drawn on previous components. Due to the limitations of working memory (WM), people usually store such information on external displays. Hegarty has provided evidence that diagrammatic reasoning is bounded to WM abilities. The more we proceed in inferring animation on later components, the more the inferences on earlier components degrade (see for instance Figure 2); "as more components of the system are 'read into' spatial working memory, the activation of all items is degraded, so that when later components are in, there is not enough activation of the later components to infer their motion" (Hegarty 2000, p 201).

[5] On this, I rely on the framework assumed by the cognitive-load theory (Paas et al 2010)

a



b

Cycle	Action Taken
1	Rope Strand 1 is inspected (simulation is given knowledge that this rope strand is being pulled)
2	Rope Strand 1 is animated (i.e. its direction of motion is inferred)
3	Rope Strand 2 is inspected
4	Rope Strand 2 is animated
5	Rope Strand 3 is inspected Pulley 1 is inspecte
6	Rope Strand 3 is animated Pulley 1 is animated
7	Rope Strand 4 is inspected
8	Rope Strand 4 is animated
9	Rope Strand 5 is inspected Pulley 2 is inspected
10	Rotation of Pulley 2 is inferred Translation of Pulley 2 is inferred
11	Weight is inspected
12	Weight is animated

Figure 2. (a) Example of diagram of a simple pulley system that can be mentally animated (b) Description of typical actions that can be one by one to animate the pulley system. Both figures taken from Hergaty (2000)

The actual limit of our cognitive architecture on this respect may be debated, and it is an empirical issue. The important point is that *no matter our external displays*, for very large systems (such as Figure 3) it is very unlikely that human cognition will be able to process all information about elements interactivity. This is because by animating components one-by-one, even if we use sophisticated instruments such computer simulations, still inferences on earlier components will degrade. This means that build-it tests will be very ineffective, if not impossible. In terms of narratives, recipes and mechanistic models, this means that for large mechanistic diagrams with many model components, no human would be able to anticipate the consequences of modifying a step in the model for all the other steps of the model, even if a computer simulation shows that the phenomenon can be possibly produced by the complex model. The computer simulation may highlight certain aspects (as Bechtel in 2016 notes), but the model is not intelligible in the sense required by mechanistic philosophy. *If the model is not intelligible in this way, then it cannot be possibly turned into an explanation*.

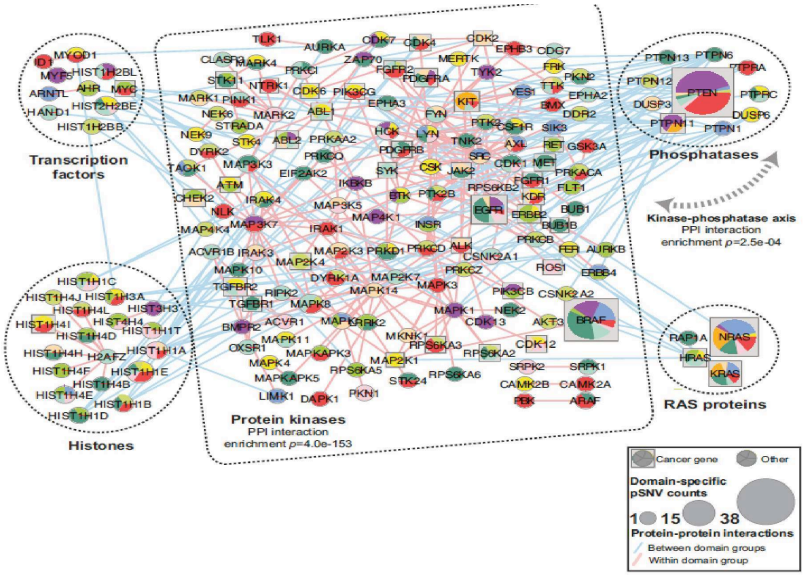


Figure 3. Network of interactions of proteins with significant enrichment of phosphorylation-related single nucleotide variations. Phosphorylation is a central post-translational modification in cancer biology. Authors are not trying to recompose the mechanism that from phosphorylated proteins (nodes) lead to a tumor phenotype, but rather to identify the magnitude of the impact of this process on cancer genes. Figure taken from (Reimand et al 2013)

The results of Hegarty's research suggest that when mechanistic models are concerned, strategies of localization are effective (in terms of explanatory potential) only when a limited number of model components are actually identified. The number may increase if we use computer simulations. However, for very large amounts of model components (such as Figure 3) recomposition is just impossible for humans, because inferences on the role of components in the causal division of labor of a phenomenon will degrade to make place for inferences about other components. This of course holds only if we have explanatory aims in mind.

To summarize, in section 2 we have made three claims:

1. If a how-possibly model can be turned into an explanation, then it is intelligible
2. If a model is not intelligible, then it cannot become explanatory
3. Complex models are a class of non-intelligible models

3. MACHINE LEARNING AND LOCALIZATIONS

Machine learning (ML) is a subfield of computer science which studies the design of computing machinery that improves its performance as it learns from its environment. A ML algorithm extracts knowledge from the input data, so that it can give better solutions to the problem that it is meant to solve. This learning process usually involves the automatic construction and refinement of a model of the incoming data. In ML terminology, a model is an information structure which is stored in the computer memory and manipulated by the algorithm.

As mentioned before, the concept of 'problem' in ML has a specific meaning which is different from other fields of science. A ML problem is defined by a set of input variables, a set of output variables, and a collection of samples which are input-output pairs. Solving a problem here means finding a quantitative relation between inputs and outputs in the form of a predictive model, in the sense that the algorithm will be used to produce a certain output given the presence of a specific input.

3.1 The PARADIGM algorithm

ML has been applied in the molecular sciences in many ways (Libbrecht and Noble 2015). Especially in cancer research[6], computer scientists have created and trained a great deal of algorithms in order to identify entities that are likely to be involved in the development of tumors, how they interact, to predict phenotypes, to recognize crucial sequences, etc (see for instance Leung et al 2016).

As a topical example of ML applied to biology, we introduce an algorithm called PARADIGM (Vaske et al 2010). This algorithm is used to infer how genetic changes in a patient influence or disrupt important genetic pathways underlying cancer progression. This is important because there is empirical evidence that "when patients harbor genomic alterations or aberrant expression in different genes, these genes often participate in a common pathway" (Vaske et al 2010, p i237). Because pathways are so large and biologists cannot hold in their mind the entities participating in them, PARADIGM integrate several genomic datasets – including datasets about interactions between genes and phenotypic consequences – to infer molecular pathways altered in patients; it predicts

[6] See for instance The Cancer Genome Atlas at https://cancergenome.nih.gov

whether a patient will have specific pathways disrupted given his/her genetic mutations.

The algorithm is based on a simplified model of the cell. Each biological pathway is modeled by a graph. Each graph contains a set of nodes, such that each node represents a cell entity, like a mRNA, a gene or a complex. A node can be only in three states (i.e. activated, normal or deactivated). The connections among nodes are called factors, and they represent the influence of some entities on other entities. It must be noticed that the model does not represent why or how these influences are exerted. Only the sign of the influence, i.e. positive or negative, is specified.

The model specifies how the expected state of an entity must be estimated. The entities which are connected by positive or negative factors to the entity at hand cast votes which are computed by multiplying +1 or –1 by the states of those entities, respectively. In addition to this, there are 'maximum' and 'minimum' connections to cast votes which are the maximum or the minimum of the states of the connected entities, respectively. Overall, the expected state of an entity is computed as the result of combining several votes obtained from the entities which are connected to it. Such a voting procedure can be associated to localizations (i.e. whether a node is activated or not), but hardly to biological explanations.

The states of the entities can be hidden, i.e. they can not be directly measured on the patients, or observable. The states of the hidden variables must be estimated by a probabilistic inference algorithm, which takes into account the states of the observed variables and the factors to estimate the most likely values of the hidden variables. Here it must be pointed out that this algorithm does not yield any explanation about the computed estimation. Moreover, it could be the case that the estimated values are not the most likely ones, since the algorithm does not guarantee that it finds the globally optimum solution.

The size of the model is determined by the number of entities and factors that the scientist wishes to insert. A larger model provides a perspective of the cell processes which contains more elements, and it might yield better predictions. This means that the more components the model has, the better the algorithm will perform. In biological terms, the larger the model, the more precise *complex localizations* the algorithm will identify, in particular by pointing more precisely towards pathways that are likely to be

disrupted in the patient with more information about the state of gene activities, complexes and cellular processes. Importantly, PARADIGM does not infer new genetic interactions, but it just helps identifying those known interaction in a new data set. It is completely supervised, in the sense that "[w]hile it infers hidden quantities (…), it makes no attempt to infer new interactions not already present in an NCI [National Cancer institute database] pathway" (Vaske et al 2010, p i244).

4 COMPLEX MODELS AND MECHANISTIC EXPLANATIONS

Before unwinding our conclusions, let me recall the results of Section 2 very briefly:

1. If a how-possibly model can be turned into an explanation, then it is intelligible[7].
2. If a model is not intelligible, then it cannot become explanatory
3. Complex model (in the sense explained in 2.2) are not intelligible

What does this have to do with PARADIGM? It is important to emphasize what we have pointed out in Section 3.1, namely that an algorithm like PARADIGM is more efficient when working with more components. If we think about models generated by algorithms such as PARADIGM in mechanistic terms, this means that the algorithm provides more precise complex localizations, because more entities that are likely to be causally relevant to a phenomenon are identified, and the information about the probability of a pathway being disrupted in a patient will be more precise. However, the models will be more complex, and they will be decreasingly intelligible. This is because the final model will count an elevated number of components, and recomposing these components into a full-fledged mechanistic explanation of how a tumor is behaving will be cognitively very difficult; the inferences about the behavior of components are not run in parallel, but one by one, and once we proceed in inferring the behavior of a component on the basis of the behavior of another component, other inferences will degrade, as Hegarty's studies have shown. In the ideal situation, PARADIGM will generate unintelligible models:

[7] Remember: A mechanistic model *x* is intelligible to a modeler *y* if *y* can use the information about the components of *x* to instantiate so-called 'build-it test'. Such tests are performed on how-possibly models to turn them into explanatory models by obtaining information on how to recompose a phenomenon (i.e. by showing how a list of biological entities are organized to produce a phenomenon).

4. Algorithms such as PARADIGM generate models which are not intelligible because such models are too complex
5. Because of 2, 3 and 4, complex models generated out of algorithms like PARADIGM cannot become explanations

This means that when we use algorithms such as PARADIGM to cope with the complexity of biological systems, we successfully handle big data sets, but such a mastery comes at a price. Using ML in molecular biology means providing more detailed localizations, but we also lose explanatory power, because no modeler will be able to recompose the mechanism out of a long list of entities.

This implies that, in the mechanistic epistemic horizon, the central role assigned to explanations should be reconsidered when contemporary molecular biosciences are concerned. As Bechtel has also emphasized in the context of computational models in mechanistic research (2016), such tools are useful to show whether some entities are likely to be involved in a particular phenomenon or suggest alternative hypotheses about the relation between certain entities. However, providing fully-fledged mechanistic explanations is another thing. It is the same with algorithms of ML; we identify more entities likely to be involved in a mechanism, we may even find out that entities involved in specific process may be connected with entities involved in other processes (via for instance Gene Ontology enrichments), but we cannot recompose a mechanism out of a list of hundreds of entities. In fact, we come to value different epistemic values, and *explanatory power is not one of them*. This somehow implies also a shift in the way scientific articles are organized; if in 'traditional' molecular biology evidence converges towards the characterization of a single mechanism, in data-intensive biology we make a list of entities that can be involved in a phenomenon, but we do not necessarily connect those entities mechanistically (Alberts 2012). Another strategy (Krogan et al 2015) – though motivated more by biologically rather than cognitive reasons – is to abstract from macromolecular entities and consider only aggregates of them in the form of networks; whether establishing network topology is providing a mechanistic explanation remains an open question.

REFERENCES

Alberts, B. (2012). The End of "Small Science"? *Science*, *337*(September), 1230529.

Bechtel, W. (2016). Using computational models to discover and understand mechanisms. *Studies in History and Philosophy of Science Part A*, *56*, 113–121.

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity - Decomposition and Localization as Strategies in Scientific Research*. Cambridge, Massachusetts, and London, England: The MIT Press.

Craver, C. (2007). *Explaining the Brain - Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. Chicago: The University of Chicago Press.

De Regt, H. (2017). *Understanding Scientific Understanding*. Oxford: Oxford University Press.

de Regt, H. W. (2015). Scientific understanding: truth or dare? *Synthese*, *192*(12), 3781–3797. http://doi.org/10.1007/s11229-014-0538-7

Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese*, *172*(2), 269–281.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford University Press.

Hegarty, M. (2000). Capacity Limits in Mechanical Reasoning. In M. Anderson, P. Cheng, & V. Haarslev (Eds.), *Diagrams 2000* (pp. 194–206). Springer-Verlag.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A., & Ideker, T. (2015). The Cancer Cell Map Initiative: Defining the Hallmark Networks of Cancer. *Molecular Cell*, *58*(4), 690–698.

Levy, A. (2014). What was Hodgkin and Huxley's achievement? *British Journal for the Philosophy of Science*, *65*(3), 469–492.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about Mechanisms. *Philosophy of Science*, (67), 1–25.

Morrison, M., & Morgan, M. (1999). Models as mediating instruments. In M. Morrison & M. Morgan (Eds.), *Models as Mediators*. Cambridge University Press.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: The MIT Press.

Ratti, E. (2018). "Models of" and "models for": On the relation between mechanistic models and experimental strategies in molecular biology. *British Journal for the Philosophy of Science*.

Reimand, J., Wagih, O., & Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Scientific Reports*, *3*.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., … Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, *26*(12), 237–245.

Vasu, K., & Nagaraja, V. (2013). Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiology and Molecular Biology Reviews*, *77*(1), 53–72.

The Roles of Possibility and Mechanism in Narrative Explanation

Abstract

There is a fairly longstanding distinction between what are called the *ideographic* as opposed to *nomothetic* sciences. The nomothetic sciences, such as physics, offer explanations in terms of the laws and regular operations of nature. The ideographic sciences, such as natural history (or, more controversially, evolutionary biology), cast explanations in terms of narratives. This paper offers an account of what is involved in offering an explanatory narrative in the historical (ideographic) sciences. I argue that narrative explanations involve two chief components: a possibility space and an explanatory causal mechanism. The presence of a possibility space is a consequence of the fact that the presently available evidence underdetermines the true historical sequence from an epistemic perspective. But the addition of an explanatory causal mechanism gives us a reason to favor one causal history over another; that is, causal mechanisms enhance our epistemic position in the face of widespread underdetermination. This is in contrast to some recent work that has argued against the use of mechanisms in some narrative contexts. Indeed, I argue that an adequate causal mechanism is always involved in narrative explanation, or else we do not have an explanation at all.

1. Introduction

The historical sciences (geology, paleontology, evolutionary biology, etc.)[1] are usually thought to deploy different explanatory strategies than the non-historical sciences (Turner 2007; Turner 2013). Whereas physics, say, seeks explanations given in terms of general laws and the like, the historical sciences seek to explain in terms of narratives. In this paper I will argue for a version of narrative explanation involving two chief components: possibility spaces and causal mechanisms. It has recently been argued that complex historical narratives (to be defined later) can't support explanations involving causal mechanisms (Currie 2014). I argue that this is mistaken. I'll go over some recent work on the history of abiogenesis research to support this contention.

The argument presented in this paper will defend two primary claims: (1) the conceptual structure of narrative explanations nearly always involves a space of alternative possibilities. This can be for either epistemic or ontological reasons. From and epistemic perspective possibility spaces are necessary on account of our position relative to the available evidence. That is, the available evidence radically underdetermines any particular causal history, and on the basis of that fact many possible histories appear compatible with what we know (see Gordon and Olson 1994, p. 15). Construed ontologically, a set of historical facts might involve a high degree of objective contingency—it might be the case that things really could have gone a number of different ways. For the purposes of this paper I remain silent with respect to this ontological aspect and defend the importance of possibility spaces for largely epistemic reasons. (2) Adequate causal mechanisms enhance our epistemic position relative to alternative causal

[1] I note that the idea of evolutionary biology as a properly "historical science" is a controversial one. See Ereshefsky (1992) for some strong arguments against the idea of evolutionary biology as having a distinctively 'historical' flavor.

histories. Causal mechanisms put us in a position to better assess the plausibility of a given history within our possibility space, and in this way enhance the epistemic power of a purportedly explanatory historical narrative. This can involve either the actual discovery of such mechanisms, or raw theoretical innovation. Citing an adequate causal mechanism may not discriminate between possibilities in decisive fashion. Rampant underdetermination seems to rule out such a possibility (see Turner 2007). But an adequate mechanism does make a given explanation more explanatory than its competitors, and so part of the task is to see how this notion of mechanistic adequacy can be cashed out in such a way as to make this notion of *explanatoriness* epistemically significant and not simply *ad hoc*.

2. The Role of Possibility Spaces

In the introduction I said that I would defend two major claims: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. This section will address the first claim by giving a more detailed account of the conceptual structure of narrative explanations and why the role of possibility spaces is so central to them.

When confronted with a natural historical problem (e.g. accounting for the processes involved in the formation of atoll reefs, say (see Ghiselin 1969)) it is my claim that what we are confronted with is, in fact, a space of *possible* histories. That is, when the historical scientist attempts to answer the question, "What geological process accounts for the formation of atoll reefs?" she understands—perhaps implicitly—that there are a number of ways things *might* have gone: she sees many possible histories. This space of possible histories essentially generates a contrasting set of possible explanations, each possible history corresponding roughly to one

hypothetical solution to the problem.[2] Obviously there's just one causal history that actually obtained, but the evidential situation is such that this history is not uniquely fixed from an epistemic perspective (see Roth 2017). The historical scientist's explanatory task then consists in finding the best approximation of the true causal history.

A nice example of this sort of reasoning process can be glimpsed in the debates over speciation processes among evolutionary biologists and paleobiologists. Stephen J. Gould and Niles Eldredge (1972) developed the theory of *punctuated equilibria* to account for the pattern of speciation witnessed in the fossil record. The idea of punctuated equilibria, in brief, holds that evolutionary change occurs in sudden bursts (on geological timescales, anyway), followed by long periods of relative evolutionary stasis. The going theory of evolutionary change at the time held to *phyletic gradualism*—the idea that the pace of evolution is slow and relatively uniform (see Turner 2011). Each of these alternatives is broadly consistent with the available fossil evidence. Phyletic gradualism takes the view that the evolutionary process is gradual, and that the fossil record is very patchy. The putatively patchy character of the fossil record means that we shouldn't expect to be able to use it as a tool for faithfully reading off patterns of speciation in the actual history of life. The theory of punctuated equilibria has it that the fossil record is relatively faithful to evolutionary history, meaning that the fossil record *does* have some explanatory import with respect to uncovering important evolutionary patterns (like speciation). The evidence in the fossil record can support either interpretation.

Consider another example, this time from geology. 19th century geologists were confronted with a fascinating geological puzzle involving what were called 'erratic blocks'.

[2] I'm certainly *not* claiming that the historical scientist is in a position to generate or realize all possible histories, as the number of such alternatives is plausibly infinite. But certainly it's possible to generate quite a few, and it seems that in fact we usually do.

These hulking slabs of (usually) granite are found miles away from any related rocks, and so the obvious question to be answered is, "How did such a large piece of granite come to be deposited here?" In 1820s Europe the answer was not immediately obvious. One well-documented case involved a granite erratic in Switzerland, which was determined to be composed of primary rocks of Alpine origin, but resting on a limestone formation many hundreds of miles from any mountains (see Rudwick 2014, pp. 117-25). Several explanations were offered: that it was deposited by the waters of the Noahic deluge; that it was carried and deposited by waters traveling down the Alps from a broken mountain dam; and only later that it was carried by glacial ice and then deposited after a subsequent melt. The process of adjudicating between each such purportedly explanatory histories (whether evolutionary patterns or seemingly bizarre geological deposits) is the subject of the next section.

It's important to stress that the evidential underdetermination of historical hypotheses is quite different than underdetermination in science more broadly. Turner (2007) argues convincingly that the problem of underdetermination is rather severe in the historical sciences given that natural processes actively destroy the evidential traces on which historical scientists rely.[3] There are two points that make this worthy of note. First, it is precisely for this reason that the explanatory task of the historical scientist *necessarily* involves the generation of a possibility space. If we can think of a natural history as a story concerning the artifacts of the natural world, then what the world presents us with is a story that's missing a great many pages. The unfortunate fact of the matter is that there are many ways of filling those pages in, each of which

[3] Turner appeals to the role played by background theories in the historical sciences to motivate his point. Here, there relevant theory is *taphonomy*, which describes the mechanisms by which the relevant evidence is destroyed (remineralization, decomposition, etc.).

is broadly compatible with our evidential situation.[4] Second, widespread underdetermination is what motivates the earlier insight that the explanatory aspiration of historical science is to give the best *approximation* of the true causal history. It is implausible to think that any of the historical hypotheses we generate will fill in the missing pages perfectly, but we can have reasons to think that some hypotheses outperform others (of which more to come).

To summarize, possibility spaces are ineliminable from narrative explanations because of our epistemic position relative to the evidence at hand. What we want is to develop a causal history that explains the phenomenon in question (e.g. erratic blocks and evolutionary patterns), but right away we realize that many different and mutually incompatible histories could—hypothetically—do the trick. The construction of a space of live possibilities allows us to have some degree of confidence that we've explored the relevant alternatives.[5] Once we've developed a space of possibilities, the initial question (such as, "What accounts for the formation of atoll reefs?") becomes importantly *contrastive*: "Why *x* and not *x'*?" where *x* and *x'* are alternative possible causal histories accounting for the target phenomenon. We want to know how it is that possibilities come to be "foreclosed" upon as a narrative explanation develops, as Beatty (2016) puts it.

3. Causal Mechanisms and Hypothesis Adjudication

[4] See Turner (2011) chapter 2 for more in-depth discussion.

[5] There's a way of reading this that might tempt one to see this as something akin to *inference to the best explanation*. Any such connection is largely superficial. The primary reason for this is that the explanatory scheme that I'm outlining is not meant to be making any especially strong claims about the strength of an explanation as related to its connection to reality. Perhaps none of the causal histories we generate are very accurate as descriptions of the true causal history.

I now turn my attention to an explication and defense of (2): adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. Causal mechanisms are what provide reasons for preferring one possible causal history over another as regards the space of possible histories generated by the natural historical problem at hand.

3.1. Mechanistic set-ups-

Because contingency is generally seen as playing such a fundamental role in natural historical contexts, the relevant mechanisms are not likely to be cashed out in terms of 'invariances' and 'regularities,' as is common in other scientific contexts (see Havstad 2011; Darden and Craver 2002). For the purposes of natural history we might instead think in terms of a more minimal conception of causal mechanisms that I'll call *mechanistic set-ups*. A mechanistic set-up differs from paradigmatic mechanisms (as in Glennan (2002))[6] in that it will often be the case that mechanistic set-ups are the result of one-off circumstances. Paradigmatic mechanisms characterize causal systems that are largely stable across time (think of protein synthesis, for instance). Mechanistic set-ups are not stable across time in this way, but still render outcomes causally expectable given that the right antecedent conditions obtain. That is, given that the right antecedent conditions obtain (and this may, of course, be a *highly contingent* affair), the causal output of the system is fully determined—we have a case of mechanical causal output.

Nancy Cartwright and John Pemberton (2013) give a simple example of a mechanistic set-up using a toy sailboat. When the toy boat is placed in the water it displaces enough liquid to

[6] "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations."

stay afloat; it has a wind-catching device for locomotion; the wind-catching device is acted about by wind gusts in order to achieve locomotive action. If we take this example as having to do with the actions of an *agent* that brings about the mechanistic set-up then we might incline toward an interpretation of the situation in terms of paradigmatic mechanisms. But imagine there's no agent involved at all; that is, let it be the case that nobody placed the boat on the water, and likewise nobody chose any windy day in particular for the use of the boat. Instead suppose that it is a series of contingent events (a child threw the boat in the garbage, it fell out of the garbage truck on the highway, and is now on the surface of a local pond, etc.) that have made things such that the boat is at some later time moving across the top of the water in the expected way.

The one-offness of the circumstances in the revised toy boat example doesn't seem to make the situation non-mechanistic in character. Rather, the mechanism just isn't stable across time in the same way paradigmatic mechanisms are. This is a mechanism in a more minimal sense: it is a mechanistic set-up. In other words, the realization of appropriate antecedent conditions renders the outcome causally expectable, even though the antecedent conditions are highly contingent.[7]

This case is so simple that it won't have much bite against Currie. Recall that Currie's claim is that mechanisms show to be of no use in *complex* narratives. In these cases the explanatory targets are *diffuse*, meaning that they involve complex networks of causal contributors (Currie 2014). An example of a diffuse target is Sauropod gigantism, Gigantism involves, at least, skeletal pneumatization, ovipary, increased basal metabolic rate, etc. Nothing seems to unify such causal contributions, and so there is no *mechanism* for gigantism, according to Currie—the explanatory target is *too diffuse* in complex narratives.

[7] See chapter 3 of Conway Morris (2003) for an in-depth discussion.

3.2. Abiogenesis, mechanistic set-ups, and hypothesis adjudication-

Abiogenesis, I argue, qualifies as a minimal mechanistic set-up in the sense just argued for. That is, the set of facts that determined the development of the very first self-replicating, heterotrophic organisms are plausibly subject to a high degree of contingency (see Conway Morris 2003), but even so, life is a deterministic consequence of just such a contingent set of facts.[8] Further, the instances that the theory aims to explain (e.g. self-replicating molecular systems; heterotrophic metabolic systems; protective membrane enclosures, etc.) are diffuse in the same sense as Sauropod gigantism. My aim here is not to give a full theoretical survey of abiogenesis, but instead to provide just enough content to justify the claim that work in this area fulfills the description of narrative already given, and that causal mechanisms play an important explanatory role, specifically to do with hypothesis adjudication.

Probably the first serious theoretical work on the origins of life is A.I. Oparin's 1923 *The Origins of Life* (Falk and Lazcano 2012). The basic theoretical framework is familiarly Darwinian. Oparin had in mind a model of biological origins whereby life comes on-line in stages, rather than all at once. The prebiotic world, on this view, was one of something approximating 'molecular competition.' For Oparin this amounted to chemical assemblages witnessing differential stability, approximately underwriting a growth model of molecular evolution (Falk and Lazcano 2012; Pigliucci 1999). The primary thing to be explained, on this model, was the development of heterotrophic metabolism. Metabolic pathways are so complex

[8] Some recent work in origins of life research may end up giving reasons to question the assumed contingency of life's emergence. See Kauffman (1993) for a classic treatment of the "self-organization" thesis, and England (2015) for more recent theoretical developments.

that Oparin thought their development must be accounted for in a basically stepwise fashion. Differential stabilities of chemical assemblages would make it such that certain molecules would make up increasingly large proportions of the chemical 'population,' making them live candidates for further downstream innovation (like complex metabolic pathways).

Oparin-type selection models have mostly—though perhaps not entirely—fallen by the wayside. Contemporary work is focused primarily on accounting for the possibility of self-replication and autocatalysis (Penny 2005). The thought is that biological origins must be accounted for in something like a two-step process, one involving the development of self-replicating material suitable for hereditary mechanisms, and another for things like metabolism and heterocatalytic functions like protein construction (Falk and Lazcano 2012; Conway Morris 2003). One of the more promising research strains in this area concerns what's known as the 'RNA World' (Conway Morris 2003). It's widely believed to be the case that the first replicators were RNA (or RNA-like) molecules. So, RNA World researchers are attempting to simulate the conditions of the prebiotic Earth in the laboratory in order to see whether the RNA model of biological origins can carry its empirical weight.

Of note for the purposes of this paper is that the dispute between metabolism-first and replication-first models of abiogenesis is precisely over whether the causal mechanisms in play can adequately account for the target phenomenon: namely, the development of living organisms in the ancient history of Earth. H.J. Muller developed a theoretical agenda stressing the need for self-replicators at the historical foundations of life (Falk and Lazcano 2012). Oparin took heterotrophic metabolic pathways as the primary puzzle to be solved (Oparin 1938; Falk and Lazcano 2012). The replication-first view has emerged as the going view among contemporary researchers primarily because it offers a more plausible mechanism for life's early development.

In order to build complex metabolic pathway it seems like it's first necessary to have a genome space that's large enough to enable downstream innovation of complex functions. So it is that the replication-first view and the research agenda dictated by projects like RNA World are taken to be more explanatory than Oparin-type explanations given in terms of selection among molecular assemblages.

4. Putting Things Together

Let's recall once more the two key claims being advanced: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. Widespread underdetermination in the historical sciences leads to the persistent appearance of possibility spaces as specified by (1), and the development of adequate causal mechanisms specified under (2) enhances our ability to adjudicate the alternatives we're faced with. Causal mechanisms put us in a position to address the contrastive question, "Why *x* and not *x'*?" Causal mechanisms are the devices by which historical counterfactuals become foreclosed upon in the sense of Beatty (2016).

Because explanation in the historical sciences is contrastive in the above sense, I argue that some notion of mechanism is involved in *every* case of successful narrative explanation. Currie (2014) argues that causal mechanisms are appropriate only for the purposes of simple narratives apt to be embedded in terms of regularities. Complex narratives with their diffuse explanatory targets require something more piecemeal that doesn't count as a causal mechanism. My more minimal conception of causal mechanisms given in terms of *mechanistic set-ups* sheds light on why this can't be right. Mechanistic set-ups aren't stable across time like paradigmatic

mechanisms, and yet we have good reason to think that the consequences of such set-ups are mechanistically determined (see Penny 2005; Glennan 2010).[9] It is just this sort of conception of mechanism that helps us to make sense of explanatory success in abiogenesis (such as it is).

Surely the genesis of the first biotic creatures is every bit as diffuse an explanatory target as Sauropod gigantism. I've argued (and I think convincingly) that it is precisely due to the adequacy of some underlying mechanism that one explanatory agenda in abiogenesis has been accepted over the alternatives. The complexity of the narrative and the diffuseness of the explanatory target appear to be beside the point. Without an adequate mechanism—however minimally construed—we can't answer the contrastive question, and so we have no explanation at all.

5. Objection and a Reply

According to Currie (2014) mechanistic set-ups (*ephemeral mechanisms* (Glennan 2010)) look like they're simply pointing to claims about sensitivity to initial conditions. If that's right, then there's a problem, because causal processes in natural historical contexts are often thought to be contingent not just in the sense that they display sensitivity to initial conditions. Such processes are taken to be subject to contingencies in a more robust sense involving "causal cascades" themselves (Currie 2014). It is not unreasonable, for instance, to think that whether a chemical assemblage will manage to hit the right configuration and produce a self-replicating RNA strand is not just a matter of realizing the right set-up conditions (independent of the chances of hitting

[9] Penny notes some interesting experimental results in which living organisms are frozen to near absolute zero, meaning that all information concerning the positions and velocities of the particles in their make-up is lost. They can, nonetheless, be successfully reanimated. Given that the only information that's retained after such a deep freezing involves the chemical structure of the organisms, a natural inference is that 'life' is a mechanical consequence of chemical parts.

on such a configuration). Whether the chemical elements enter into the appropriate causal relations for manifesting autocatalysis might *itself* be a probabilistic matter. Having the right elements might not be all you need—you might need the right elements plus a bit of probabilistic luck. Objective probabilities of this sort might do some damage to the mechanistic account, since it would seem not to be the case that an explanandum *just follows* from a causal set-up. The force of this objection is at least partly dependent on one's answer to the question of where in the world we ought to 'place' objective chances (if there are any).

Most of our intuitions about objective probabilities (probably) derive from our ongoing observations of the world. A lot of stuff in the world *just seems* chancy. We regularly speak in terms of the "odds" or "chances" of developing cancer and the like. Simplifying quite a bit, when we say that there's a 40 percent chance that Susan will live for more than 5 years after being diagnosed with some cancer that has developed to some particular stage, what we're saying is that approximately 40 percent of people that present as cases sufficiently similar to Susan have lived for 5 years or more. One way to read this is in terms of causal indeterminacy. That is, there is really no matter of the fact at time *t* as to what will be the case at time *t'*, aside from the probabilistic facts about cancer populations. The future is (to some degree) causally open, as the causal cascades are operating in a fundamentally probabilistic way.

Such a reading, however, is by no means forced. Bruce Glymour (1998) offers a picture wherein objective probabilities are placed at the level of causal *interactions*. That is, entities *e* and *e** enter into causal interactions with each other on a probabilistic basis, but when they do, the downstream effects unfold in a fully deterministic fashion. Probabilistic partitions of the world, then, are just reflections of whether certain causal interactions became manifest in certain subpopulations or not. If 40 percent of patients with a certain cancer at a particular stage will

survive for more than five year, it's because free radicals (probabilistically) failed to enter into certain causal interactions with healthy cells. The opposite is the case for the contrasting class of fatal cases. On this picture, determinism of the relevant kind seems to be preserved. In such cases as the right causal interactions are realized, downstream effects unfold in mechanical fashion.

6. Conclusion

In this paper I argued for two main claims: (1) the conceptual structure of narrative explanation nearly always involves a space of alternative possibilities, and (2) adequate causal mechanisms enhance our epistemic position relative to alternative causal histories. The reason that narrative explanations involve possibility spaces has to do with our epistemic position relative to the available evidence. Undetermination so permeates the historical sciences that any problem for which we seek an explanation will involve an array of possible alternative causal histories, each of which is broadly consistent with the available evidence. It is the introduction of an adequate causal mechanism that puts us in a position to improve our epistemic lot—with a good mechanism in hand, we can begin to foreclose upon alternatives.

References

Beatty, John. 2016. “What Are Narratives Good For?” *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 58. Elsevier Ltd: 33–40. doi:10.1016/j.shpsc.2015.12.016.

———. 2017. “Narrative Possibility and Narrative Explanation.” *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 1–14. doi:10.1016/j.shpsa.2017.03.001.

Cartwright, Nancy, and John Pemberton. 2013. "Aristotelian Powers: Without Them, What Would Science Do?" in Groff & Greco (Eds.), *Powers and Capacities in Philosophy: The New Aristotelianism*. New York: Routledge.

Conway Morris, Simon. 2003. *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge: Cambridge University Press.

Currie, Adrian Mitchell. 2014. “Narratives, Mechanisms and Progress in Historical Science.” *Synthese* 191 (6): 1163–83. doi:10.1007/s11229-013-0317-x.

Darden, Lindley, and Carl Craver. 2002. "Strategies in the interfield discovery of the mechanism of protein synthesis." *Studies in History and Philosophy of Biological and Biomedical Sciences* 33: 1-28.

Eldredge, Niles, and Stephen J. Gould. 1972. "Punctuated equilibria: an alternative to phyletic gradualism," in Schopf (Ed.), *Models in Paleobiology*. San Francisco: Freeman Cooper.

England, Jeremy. 2015. "Dissipative Adaptation in Self-Driven Assembly." *Nature Nanotechnology*, 10: 919-923.

Ereshefsky, Marc. 1992. "The Historical Nature of Evolutionary Theory." In *History and Evolution*, ed. Matthew Nitecki and Doris Nitecki. New York: The SUNY Press.

Falk, Raphael, and Antonio Lazcano. 2012. “The Forgotten Dispute: A.I. Oparin and H.J. Muller on the Origin of Life.” *History and Philosophy of the Life Sciences* 34 (3): 373–90.

Ghiselin, Michael T. 1969. *The Triumph of the Darwinian Method*. Chicago: Chicago University Press.

Glennan, Stuart. 1996. “Mechanisms and the Nature of Causation.” *Erkenntnis* 44 (1): 49–71.

———. 2002. “Rethinking Mechanistic Explanation.” *Philosophy of Science* 69 (S3): S342–53.

———. 2010. “Ephemeral Mechanisms and Historical Explanation.” *Erkenntnis* 72 (2): 251–66. doi:10.1007/s10670-009-9203-9.

Glymour, Bruce. 1998. “Contrastive, Non-Probabilistic Statistical Explanations.” *Philosophy of Science* 65 (3): 448–71.

Gordon, Malcolm and Everett Olson. 1994. *Invasions of the Land*. New York: Columbia University Press.

Haldane, J.B.S. 1954. "The origin of life." *New Biology* 16: 12-27.

Havstad, Joyce C. 2011. “Problems for Natural Selection as a Mechanism.” *Philosophy of Science* 78 (3): 512–23. doi:10.1086/660734.

Hull, David. 1975. “Central Subjects and Historical Narratives.” *History and Theory* 14 (3): 253–74.

Jeffares, Ben. 2008. “Testing Times: Regularities in the Historical Sciences.” *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 39 (4). Elsevier Ltd: 469–75. doi:10.1016/j.shpsc.2008.09.003.

Kauffman, Stewart. 1993. *The Origins of Order: Self Organization and Selection in Evolution*. Oxford: Oxford University Press.

Mink, Louis O. 1970. "History and Fiction as Modes of Comprehension." *New Literary History*, 1 (3): 541-558.

Oparin, A.I. 1938. *The Origin of Life*. New York: MacMillan.

Penny, David. 2005. “An Interpretive Review of the Origin of Life Research.” *Biology & Philosophy* 20 (4): 633–71. doi:10.1007/s10539-004-7342-6.

Pigliucci, Massimo. 1999. "Where do we come from? A humbling look at the biology of life's origin." *Skeptical Inquirer* 99: 193-206.

Ricoeur, Paul. 1984. *Time and Narrative (Volume 1)*. Chicago: University of Chicago Press.

Roth, Paul A. 2017. “Essentially Narrative Explanations.” *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 1–9. doi:10.1016/j.shpsa.2017.03.008.

Rudwick, M.J.S. 2014. *Earth's Deep History: How It Was Discovered and Why It Matters*. Chicago: Chicago University Press.

Sepkosi, David. 2012. *Rereading the Fossil Record: The Growth of Paleontology as an Evolutionary Discipline*. Chicago: Chicago University Press.

Sunstein, Cass R. 2016. “Historical Explanations Always Involve Counterfactual History.” *Journal of the Philosophy of History* 10 (3): 433–40. doi:10.1163/18722636-12341345.

Turner, Derek. 2013. “Historical Geology: Methodology and Metaphysics.” *Geological Society of America Special Papers* 502 (2): 11–18. doi:10.1130/2013.2502(02).

———. 2007. *Making Prehistory: Historical Science and the Scientific Realism Debate*. Cambridge: Cambridge University Press.

———. 2011. *Paleontology: A Philosophical Introduction*. Cambridge: Cambridge University Press.

A Better Foundation for Public Trust in Science

S. Andrew Schroeder
Claremont McKenna College/Princeton University
aschroeder@cmc.edu

draft of 15 June 2018

Abstract. There is a growing consensus among philosophers of science that core parts of the scientific process involve non-epistemic values. This undermines the traditional foundation for public trust in science. In this paper I consider two proposals for justifying public trust in value-laden science. According to the first, scientists can promote trust by being transparent about their value choices. On the second, trust requires that the values of a scientist align with the values of an individual member of the public. I argue that neither of these proposals work and suggest an alternative that does better: when scientists must appeal to values in the course of their research, they should appeal to *democratic values*, the values of the public or its representatives.

1. Introduction

The American public's trust in science is a complicated matter. Surveys reveal that trust in science has remained consistently high for decades, and scientists remain among the most highly-trusted professional groups (Funk 2017). However, within some segments of society (especially conservatives) trust has declined significantly (Gauchat 2012), and there are obviously serious gaps in trust on certain issues, such as climate change, vaccine safety, and GM foods (Funk 2017). The picture, then, is a complex one, but on balance it is clear that things would be better if the public placed greater trust in science and scientists, at least on certain issues.

As a philosopher, I am not in a position to determine what explains the lack of trust in science, nor to weigh on what will in fact increase trust. Instead, in this paper I will look at the question of what scientists can do to *merit* the public's trust — under what conditions the public *should* trust scientists. Indeed, it seems to me that we need to answer the normative question first: if we take steps to increase

public trust in science, our goal should not simply be to make scientists *trusted*, we should also want them to be *trustworthy*.

In what follows, I'll first explain how recent work in the philosophy of science undermines the traditional justification given to the public for trusting science. I'll then consider two proposals that have been offered to ground public trust in science: one calling for transparency about values, the second calling for an alignment of values. I'll argue that the first proposal backfires — it rationally should *decrease* trust in science — and the second is impractical. I'll then present an alternative that is imperfect, but better than the alternatives: when scientists must appeal to values in the course of their work, they should appeal to *democratic values* — roughly, the values of the public or its representatives.

2. Trust and the Value-Free Ideal

Why should the public trust scientists? The typical answer to that question points to the nature of science. Science, it is said, is about facts, and not values. It delivers us objective, verifiable truths about the world — truths not colored by political beliefs, personal values, or wishful thinking. Of course, there are scientists who inadvertently or intentionally allow ideology to influence their results. But these are instances of *bad* science. Just as we should not allow the existence of incompetent or corrupt carpenters to undermine our trust in carpentry, we should not allow the existence of incompetent or corrupt scientists to undermine our trust in science. So long as we have institutions in place to credential good scientists and root out corrupt ones, we should trust the conclusions of science.

There is, unfortunately, one problem with this story: science isn't actually like that. In the past few decades, philosophers of science have shown that even good science requires non-epistemic value judgments. Without wading into the nuanced differences between views, I think it is fair to say that there is a consensus among philosophers of science that non-epistemic values can appropriately play a role in at

least some of the following choices: selecting scientific models, evaluating evidence, structuring quantitative measures, defining concepts, and preparing information for presentation to non-experts.[1]

These value choices can have a significant impact on the outcome of scientific studies. Consider, for example, the influential Global Burden of Disease Study (GBD). In its first major release it described itself as aiming to "decouple epidemiological assessment from advocacy" (Murray and Lopez 1996, 247). In the summary of their ten volume report, the authors describe their study as making "a number of startling individual observations" about global health, the first of which was that, "[t]he burdens of mental illnesses…have been seriously underestimated by traditional approaches… [P]sychiatric conditions are responsible…for almost 11 per cent of disease burden worldwide" (Murray and Lopez 1996, 3). Many others have cited and relied on the GBD's conclusions concerning the magnitude of mental illness globally (Prince *et al*. 2007). And nearly two decades later, the same GBD authors, in commenting on the legacy of the 1996 study, proudly noted that it "brought global, regional, and local attention to the burden of mental health" (Murray *et al*. 2012, 3).

It turns out, however, that the reported burden of mental health was driven largely by two value choices: the choice to "discount" and to "age-weight" the health losses measured by the study. Discounting is the standard economic practice of counting benefits farther in the future as being of lesser value compared to otherwise similar benefits in the present, and age-weighting involves giving health losses in the middle years of life greater weight than otherwise similar health losses among infants or the elderly. Further details about discounting and age-weighting aren't relevant to this paper; all we need to note is that the study authors acknowledged that each reflects value judgments, and that a reasonable case could be made to omit them (Murray 1996; Murray *et al*. 2012).[2] Given other methodological choices made by the authors, these two weighting functions combine to give relatively more weight to health

[1] On these points see e.g. Reiss (2017) and Elliott (2011).

[2] Indeed, in 2012 the GBD ceased age-weighting and discounting. There was also a third value choice that drove the large burden attributed to mental health: the choice to attribute all suicides to depression (Murray and Lopez 1996, 250). Because I do not know precisely how this affected the results, I set it aside here. For much more on discounting, age-weighting, and other value choices in the GBD, see Schroeder (2017).

conditions which (1) commonly affect adults or older children (rather than the elderly or young children), (2) have disability (rather than death) as their primary impact, and (3) have their negative effects relatively close to the onset or diagnosis of the condition (rather than far in the future). It should not be surprising, then, that when the GBD authors ran a sensitivity analysis to see how the decision to discount and age-weight affected the results, they discovered that the conditions most affected by these choices — unipolar major depression, anaemia, alcohol use, bipolar disorder, obsessive-compulsive disorder, chlamydia, drug use, panic disorder, post-traumatic stress disorder — were largely composed of mental health conditions (Murray and Lopez 1996, 282). Overall, the global burden of disease attributable to psychiatric conditions drops from 10.5% to 5.6%, when the results are not age-weighted or discounted (Murray and Lopez 1996, 261, 281).

I don't want to comment here on the wisdom of the GBD scientists' decision to discount and age-weight.[3] They offer clear arguments in favor of doing so and many other studies have done the same, so at minimum I think their choices were defensible. The point is that what was arguably the top-billed result of a major study — a result which was picked up on by many others, and which was still being proudly touted by the study authors years later — was not directly implied by the underlying facts. It was driven by a pair of value judgments. Had the GBD scientists had different views on the values connected to discounting and age-weighting, they would have reported very different conclusions concerning the global impact of mental illness.[4]

This case is not unique. The dramatically different assessments given by Stern and Nordhaus on the urgency of acting to address climate change can largely be traced to the way each valued the present versus the future (Weisbach and Sunstein 2009). Similar conclusions are plausible concerning the value choices involved in classifying instances of sexual misbehavior in research on sexual assault, the value

[3] I do so in Schroeder (unpublished-a).

[4] Although the sensitivity analysis was conducted by the original study authors, they do not draw any connection to their prominent claims concerning the global extent of mental illness. To my knowledge, this paper is the first to do so.

choices impacting the modeling of low-level exposures to toxins (Elliott 2011), and the value choices involved in constructing price indices (Reiss 2008).

A natural — and not implausible — response to these cases is to suggest they are outliers. Although some scientific conclusions are sensitive to value choices, the vast majority are not. The Earth really is getting warmer and sea levels really are rising, due to human activity. Vaccines really do prevent measles and really don't cause autism. These conclusions are not sensitive in any reasonable way to non-epistemic value judgments made by scientists in the course of their research. The problem, however, is that there is no clear way for a non-expert to verify this — to tell which cases are the outliers and which are not. This, I think, justifies a certain amount of skepticism. "Although some of our conclusions do depend on value judgments, trust us that *this* one doesn't," isn't nearly as confidence-inspiring as, "Our conclusions depend only on facts, not values."

I conclude, then, that rejecting the view of science as value-free, combined with high-profile examples of scientific conclusions that do crucially depend on value judgments, undermines the claim of science to public trust in a significant way. In other words, it explains why it may be rational for the public to place less trust in the conclusions of science on a broad range of issues — including in areas, such as climate change and vaccine safety, where major conclusions are not in fact sensitive to different value judgments.[5]

3. Grounding Trust in Transparency

Good science is not value-free, which undermines the standard justification given for trust in science. What, then, can scientists do to merit the public's trust? The standard response has been to appeal to transparency. If values cannot or should not be eliminated from the scientific process, scientists

[5] For similar conclusions see Douglas (2017); Wilholt (2013); Irzik and Kurtulmus (*forthcoming*); and Elliott and Resnik (2014).

should be "as transparent as possible about the ways in which interests and values may influence their work" (Elliott and Resnik 2014, 649; *cf.* Ashford 1998; Douglas 2008; McKaughan and Elliott 2018). Obviously, in order for this proposal to work, scientists would need to be aware — much more aware than most are today — of the ways in which value judgments influence their work. But, since we have independent reason to want such awareness, let us assume that calls for transparency are accompanied by a mechanism for increasing such awareness by scientists.

Would such a proposal work? Transparency about values can help ground trust in some situations, but I see no reason to think that it should broadly support public trust in science. Transparency is only useful in supporting — as opposed to eroding — trust if it enables the recipient of that information to determine how it has affected the author's conclusions. (Knowing I have a conflict of interest will typically reduce your trust in what I tell you, unless you can determine how that conflict influenced my conclusions.) Transparency, then, will only promote trust in a robust way if the public understands how value choice influenced the results, and understands what alternative value choices could have been made and how they would have influenced the results. These criteria may be satisfiable when the effect of a value choice is relatively simple. Suppose, for example, that a scientist classifies non-consensual kissing as "sexual assault", rather than "sexual misconduct", on the grounds that she believes it has more in common with rape (a clear instance of sexual assault) than it does with contributing to a sexualized workplace (a clear instance of sexual misconduct). The value judgment here is relatively simple to explain, an alternative classification is obvious, and (if the statistics involved are simple) the effect of alternative classification on the study may be relatively straightforward. So transparency could work here.

Many value choices, however, are much more complex. Think about choices embedded in complex statistical calculations — for example, those involved in aggregating climate models (Winsberg 2012) or in calculating price indices (Reiss 2008). In cases like these, it will be very hard to clearly explain the importance of any individual value choice and harder still to explain what alternative choices

could have been made. Further, many studies involve a large number of value judgments. Schroeder (2017), for example, identifies more than ten value choices which non-trivially influenced the Global Burden of Disease Study's results. Even if each of those value choices could be explained individually, it would be virtually impossible for a non-expert to figure out the interaction effects between them.

What these cases show is that even if scientists make a serious effort at transparency — not simply listing their value judgments, but attempting to explain how those judgments have influenced their results — in many cases it simply won't be possible to communicate to the public how those values have impacted their work.[6] And, if the public can't trace the impact of those values, transparency doesn't amount to much more than a warning — a reason to *distrust*, rather than to trust. A parallel realization can be seen in the way many medical schools and journals have handled researchers' conflicts of interest. Whereas in the past disclosures of conflicts of interest — essentially, transparency — were often regarded as sufficient; many have now realized that merely knowing about such conflicts does not appreciably help a reader to interpret a study. There is thus a growing move towards banning all significant conflicts of interest.[7]

4. Grounding Trust in an Alignment of Values

The previous section argued that transparency about values is not typically a solution to the problem of public trust in science. That problem, we can now see, was not caused by the fact that values were *hidden*; it was caused by the fact that the values of scientists may *diverge* from the values of any

[6] McKaughan and Elliott (2018, and in other works) suggest that scientists, through a particular sort of transparency, seek to promote "backtracking" — that is, to enable non-experts to understand how values have influenced scientists' results and to see how those results might have looked given alternative values. They seem to suggest that, at least in the cases they consider, this will frequently be possible. I am claiming that this will not generally be feasible. See Schroeder (unpublished-a) for a more detailed discussion of a particular case.

[7] See e.g. <https://ari.hms.harvard.edu/interim-policy-statement-conflicts-interest-and-commitment>

individual member of the public.[8] To promote public trust in science, then, it seems that we need to eliminate that divergence. This is the insight that motivates Irzik and Kurtulmus (*forthcoming*; *cf*. Douglas 2017; Wilholt 2013), who argue that what they call "enhanced" trust requires that a member of the public knows that a scientist has worked from value choices that are in line with her own.

If this proposal were feasible, I think it would provide a good foundation for trust. And, in certain limited cases, it may be feasible. When science is conducted by explicitly ideological organizations, members of the public may be able to make quick and generally accurate judgments about what values scientists hold, and accordingly may be able to seek out research done by scientists who share their values. (A pragmatic environmentalist, for example, might be confident that scientists employed by the Environmental Defense Fund are likely to share her values.)

Most science, however, is not conducted by explicitly ideological organizations. In these cases, it will typically be very hard for members of the public to confidently determine whether a given study relied on value judgments similar to her own. Even when this can be done (perhaps as a result of admirable transparency and clarity on the part of a scientist), it will require sustained and detailed engagement from the public, who will have to pay close attention not just to the conclusions of scientific studies, but also to their methodology. Although such close attention to the details of science would be beneficial for a great many reasons, it unfortunately is not realistic on a broad scale. There are simply too many scientific studies out there that are potentially relevant to an individual's decisions for even attentive members of the public to keep up. If our model for trust in science requires an alignment of values between the scientist and individual members of the public, trust in science can't be a broad phenomenon. Further, I don't think we want our foundation for trust in science to make that trust accessible only to those with the education and time to invest in exploring the details of individual scientific studies.

[8] It seems relevant to note here that distrust in science is greatest among those who identify as politically conservative, while studies show that university scientists in the U.S. overwhelming support liberal candidates for political office. Whether or not this in fact explains the distrust conservatives have in science, the argument thus far shows why such distrust could have a rational foundation.

I also — somewhat speculatively — worry that adopting this proposal would exacerbate another problem. Suppose the proposal works and, at least on some issues, members of the public are able to identify and rely upon science conducted in accordance with their own values. This, I think, might lead to a further "politicization" of science, as each side on some issue seeks scientists who share their values. Of course, once we allow a role for values in science, value-based scientific disagreement isn't necessarily a problem. Faced, for example, with one experimental design that is more prone to false positives and another that is more prone to false negatives, either choice may be scientifically legitimate. It may therefore be appropriate for more environmentally-minded citizens to rely on different studies than citizens more concerned about economic development. I worry, though, that in a culture where the public specifically seeks science done by those who share their values, it will be too easy to write off any differences in conclusions as due to value judgments — too easy for environmentalists to assume that any time pro-environment and pro-industry scientists reach different conclusions, it must be due to different underlying, legitimate value judgments. In reality, though, most such disagreements are the result of *bad* science. The worry, then, is that if we grow too comfortable with each side of an issue having its own science, it will be harder to distinguish scientific disagreements that can be traced to legitimate value judgments, from disagreements that are based on illegitimate value judgments or simple scientific error. This would be a major loss.

5. Grounding Trust in Democratic Values

I've argued that neither transparency about values nor an alignment of values can provide a broad foundation for public trust in science. Let me, then, suggest a proposal that, though imperfect, can do better. From what's been said so far, we can note a few features that a better solution should have. First, both the transparency and aligned values proposals ran into trouble because they require a great deal of attention and sophistication from the public. Most individuals simply don't have the training to

understand more technical value choices, or value choices embedded within complex calculations. And, even when such understanding is possible, it will often require a level of attention that will in practice be accessible only to the well-off. We should therefore look for a foundation for public trust which doesn't require such detailed understanding of or close attention to individual scientific studies. Second, I suggested that the aligned values proposal, in telling individuals to seek out studies conducted in accordance with their own values, could reinforce a kind of politicization that may have bad consequences. It would be better to find a proposal that wouldn't so easily divide scientists and the public along ideological lines. Third, the problem with the transparency proposal (which the aligned values proposal tried, impractically, to address) was that values, even if transparent, can be alien. In order for an individual to truly trust science, that science must be built on values that have some kind of legitimacy for her.

I think scientists can satisfy two-and-a-half of these three criteria by appealing to *democratic values* — the values of the public and its representatives — when value judgments are called for in the scientific process. The details of this proposal go beyond what I can say here.[9] But, briefly, the idea is that we look to political philosophy to tell us how to determine the (legitimate) values representative of some population. In some cases, those values might be the output of a procedure, such as a deliberative democracy exercise, a citizen science initiative, or a public referendum.[10] In other cases, it might be more appropriate to equate a population's values with the views, suitably "filtered" and "laundered", currently held by its members. ("Filtering" may be necessary to remove politically illegitimate values, e.g. racist values, and "laundering" to clean up values that are unrefined or based on false empirical beliefs.) In cases where there is a broad social consensus, that might count as the relevant democratic value; in cases where there is a bimodal distribution of values, we might say that there are two democratic values; etc.

---

[9] See Schroeder (unpublished-b) for a bit more. Many other philosophers have argued that there should be an important place for democratic values in science. See, for example, Kitcher (2011), Intemann (2015), and Douglas (2005).

[10] The extensive literature on "mini-publics" offers a promising starting point. See e.g. Escobar and Elstub (2017).

Suppose, then, that political philosophers, informed by empirical research, can give us a way of determining democratic values. I suggest that when value judgments are called for within the scientific process,[11] scientists should use democratic values when arriving at their primary or top-line results — the sort of results reported in an abstract, executive summary, or in the initial portions of the analysis. Scientists could then offer a clearly-designated alternative analysis based on another set of values, e.g. their own. I think this proposal can address two of the concerns with which I began this section, and can make some progress towards answering the third.

Let us first consider the too-much-attention and politicization problems. On the democratic values proposal, if an individual can trust that a study was competently carried out — a matter I'll return to below — then she can know, without digging into the methodological details, that its conclusions are based on objective facts plus democratic values.[12] This means that, in most cases, the public need not pay detailed attention to the methodological details of individual studies — thus solving the too-much-attention problem. Further, if scientific conclusions are based on objective facts plus democratic values, any two scientists investigating the same problem in the same social and political context should reach roughly the same conclusion. This recovers a kind of objectivity for science — not objectivity as freedom from values, but objectivity as freedom from personal biases. On this picture, the individual characteristics of a scientist should have no impact on her conclusions — a conception of objectivity that has been defended on independent grounds (Reiss and Sprenger 2014; *cf.* Daston and Galison 2007 on "mechanical objectivity"). If they are both doing good science, the environmentalist and the industrialist should reach the same top-line conclusions. And if the environmentalist and industrialist reach different

[11] This proposal is restricted to value judgments that arise within the scientific process. In particular, I do not mean for it to apply to problem selection. Scientists should be free to choose research projects that are not the projects that would be chosen by the general public. (The public, however, is under no obligation to fund such projects.) I treat the choice of research topics differently than choices that arise within the course of research because I think that scientists have different rights at stake in each case. For some related ideas, see Schroeder (2017b).

[12] There may also, of course, be methodological choices not based on non-epistemic values (including choices based on epistemic values). I set these aside here, since the problems of trust I'm concerned with don't arise in the same way from them.

top-line conclusions, it means that one or the other has made some sort of error. This, I think, provides a solution to the politicization problem: on the democratic values proposal, good science (at least in its primary analyses) will speak with a single voice.

The democratic values proposal therefore solves two of the three problems we noted above. Of course, it only does so if the public can be confident that scientists really are making use of democratic values. Why should the public assume that? Right now, I think the answer is: they shouldn't! For the democratic values proposal to work, it must be accepted by a significant portion of the scientific community, or by an easily-identifiable subset of the scientific community. If that were to happen, though, then the problem here becomes the more general one of how the public can trust scientists to enforce their own norms. The procedures and policies now in place work reasonably well, I think, to expose unethical treatment of research subjects, falsification of data, and certain other types of misconduct. I am therefore optimistic that, given a greater awareness of the role value judgments play in scientific research, a system could be devised to identify scientists who depart from a professional norm requiring the use of democratic values.

6. Science, Values, and Democracy

I've argued that the democratic values proposal can address two of the problems that faced the alternative views. But what about the third? On the transparency proposal, the values of scientists can truly be alien. If a scientist conducts research based on her own values, then, unless I happen to share those values, I have no meaningful relationship to those values. If, however, a scientists appeals to democratic values, then there is a relationship, even if I don't share those values. If democratic procedures or methods were carried out properly, then my values were an input into the process which yielded democratic values. My values are, in a sense, represented in the output of that process. This, in turn, means that those values should have a kind of legitimacy for me. In a democracy, we regularly

impose non-preferred outcomes on people when they are out-voted. So long as democratic procedures are carried out properly, this seems to be legitimate — not ideal, perhaps, but better than any available alternative. On the democratic values proposal, then, when a particular scientific conclusion is uncontested, the public can trust that that conclusion is one drawn solely from the facts, plus perhaps the values that *we* share. For most of us, who don't have the time, inclination, or ability to dig into the details of each scientific study we rely on, or who have a strong commitment to democracy, that will be enough.

I think that the foregoing provides a reasonable answer to the alien values concern. It is of course not a perfect answer. It would be better, at least from the perspective of trust, to get each member of the public access to "personalized" science conducted in accordance with her values. This, however, is impractical, as we saw when discussing the aligned values proposal. So long as that is the case, there is no way to accommodate everyone. Democratic values seem like a reasonable compromise in such a situation.

All of that said, it would be nice if we could say a bit more to those ill-served by democratic values. What should we say, for example, to an individual who knows that her values lie outside the political mainstream on some issue and is therefore distrustful of science done with democratic values on that issue? The first thing to note is that, in such cases, the democratic values proposal fares no worse (or at least not much worse) than the transparency or aligned values proposals. The democratic values proposal is fully consistent with transparency - something we have independent reason to want. So, in cases where the transparency proposal works (e.g. cases where the value choices are few, easy to understand, and computationally simple), the same advantages can be had with the democratic values proposal. Individuals who disagree with a particular value judgment and have the time and expertise to do so can determine how results would have looked under a different set of value judgments. Also, recall that I am proposing only that primary or top-line results be based on democratic values. In cases where value judgments can make a big difference — as in the Global Burden of Disease Study case discussed earlier — we might hope that scientists who hold contrary values will note the dependence of those

results on values by offering secondary, alternative analyses that begin from different value judgments. Those who have the time and expertise to dig into the methodology of scientific reports can do so, seeking out results based on values they share, as the aligned values proposal would recommend.

If the foregoing is correct, the democratic values proposal does better than the alternatives in most cases, and no worse in others. That should be sufficient reason to prefer it. But I think we can say a bit more. In what cases is the complaint from minority values most compelling? It is not, I think, when it comes from people whose values lie outside the mainstream on some issues, but within the mainstream on many other issues. The much more compelling complaint comes from people whose values consistently lie outside the mainstream — people who are consistently out-voted. Oftentimes (though of course not always) when this happens, it involves individuals who are members of groups that are or have been marginalized by mainstream society. Think, for example, of cultural or (dis)ability-based groups whose values and ways of life have been consistently treated as being less valuable and worthy of respect than the values and ways of life of the majority.

I think the democratic values proposal has two important features that can partially address such complaints. First, remember that the democratic values proposal launders and filters the actual values held by the public. Certain values — e.g. racist or sexist ones — conflict with basic democratic principles of equal worth, and so cannot be candidate democratic values. Thus, even in a racist society, telling scientists to work from democratic values will not tell them to work from racist values.[13] Second, in what I regard as its most plausible forms, democracy is not a form of government based on one person-one vote. It is a form of government based on the idea that all citizens are of equal worth and have a right to equal consideration. This suggests that, in cases where minority values are held by a group that is or has been the subject of exclusion or discrimination, democratic principles may sometimes require giving those values extra weight, or a voice disproportionate to their statistical representation in the population, as a way of accounting or compensating for past unjust treatment. Thus, democratic principles may in

[13] See Schroeder (unpublished-b) for more on this.

some cases require treating the values held by an excluded minority as democratically on a par with the conflicting values held by the majority.[14]

These considerations, I think, lessen the force of the complaint from minority values, especially in its most serious incarnation. But I don't think they eliminate it. There will still be people whose values will consistently be marginalized by the democratic view. In such cases, the main recourse available is an appeal to alternate results. If individuals with minority views can count on there being scientists who share those views, they can expect that the kind of alternative analysis they would prefer will be out there, at least in cases where it makes a difference. Of course, scientists are currently a rather homogeneous bunch along many dimensions. So this suggests that the call to work from democratic values provides (yet further) support for the importance of increasing diversity within the scientific community.[15]

[14] See Kelman (2000) for an example of this sort of argument in the context of disability.

[15] ACKNOWLEDGEMENTS TO BE ADDED

References

Ashford, Nicholas. 1988. "Science and Values in the Regulatory Process." *Statistical Science* 3.

Daston, Lorraine and Peter Galison. 2007. *Objectivity*. MIT Press.

Douglas, Heather. 2017. "Science, Values, and Citizens." In *Eppur si muove: Doing History and Philosophy of Science with Peter Machamer,* ed. Adams, Biener, Feest, and Sullivan. Dordrecht: Springer.

—. 2008. "The Role of Values in Expert Reasoning." *Public Affairs Quarterly* 22.

—. 2005. "Inserting the Public into Science." In *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*, ed. Maasen and Weingart. Dordrecht: Springer.

Elliott, Kevin. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. Oxford: Oxford University Press.

Elliott, Kevin and David Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives* 122.

Escobar, Oliver and Stephen Elstub. 2017. "Forms of Mini-Publics: an Introduction to Deliberative Innovations in Democratic Practice," NewDemocracy Research and Development Note, available at <https://www.newdemocracy.com.au/research/research-notes/399-forms-of-mini-publics>.

Funk, Cary. 2017. "Real Numbers: Mixed Messages about Public Trust in Science." *Issues in Science and Technology* 34.

Gauchat, Gordon. 2012. "Politicization of Science in the Public Sphere: A Study of Public Trust in the United States, 1974 to 2010." *American Sociological Review* 77.

Intemann, Kristin. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5.

Irzik, Gürol and Faik Kurtulmus. *Forthcoming*. "What is Epistemic Public Trust in Science?" *British Journal for Philosophy of Science*.

Kelman, Mark. 2000. "Does Disability Status Matter?" In *Americans with Disabilities: Exploring Implications of the Law for Individuals and Institutions,* eds. Francis and Silvers. Routledge

Kitcher, Philip. 2011. *Science in a Democratic Society*. Amherst, NY: Prometheus.

McKaughan, Daniel and Kevin Elliott. 2018. "Just the Facts or Expert Opinion? The Backtracking Approach to Socially Responsible Science Communication," in *Ethics and Practice in Science Communication* (eds. Priest, Goodwin, and Dahlstrom). Chicago: University of Chicago Press.

Murray, Christopher. 1996. "Rethinking DALYs." In *The Global Burden of Disease,* ed. Murray and Lopez.

Murray, Christopher and Alan Lopez (Eds). 1996. *The Global Burden of Disease*. Harvard University Press.

Murray, Christopher *et al*. 2012. Supplementary appendix to "GBD 2010: design, definitions, and metrics." *Lancet* 380.

Prince, Martin *et al*. 2007. "No health without mental health." *Lancet* 370.

Reiss, Julian. 2017. "Fact-value entanglement in positive economics." *Journal of Economic Methodology* 24.

—. 2008. *Error in Economics: The Methodology of Evidence-Based Economics*. London: Routledge.

Reiss, Julian and Jan Sprenger. 2014. "Scientific Objectivity." In *The Stanford Encyclopedia of Philosophy* (Winter 2017 edition), ed. Zalta.

Schroeder, S. Andrew. 2017. "Value Choices in Summary Measures of Population Health." *Public Health Ethics* 10.

—. 2017b. "Using Democratic Values in Science: an Objection and (Partial) Response," *Philosophy of Science* 84.

—. Unpublished-a. "Which Values Should We Build Into Economic Measures?" *Under review.*

—. Unpublished-b. "Communicating Scientific Results to Policy-makers," *manuscript on file with author.*

Weisbach, David and Cass Sunstein. 2009. "Climate Change and Discounting the Future: A Guide for the Perplexed," *Yale Law and Policy Review* 27.

Wilholt, Torsten. 2013. "Epistemic Trust in Science." *British Journal for Philosophy of Science* 64.

Winsberg, Eric. 2012. "Values and Uncertainties in the Predictions of Global Climate Models." *Kennedy Institute of Ethics Journal* 22.

PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association

Inferential power, formalisms, and scientific models

Ardourel Vincent*, Anouk Barberousse†, Cyrille Imbert§

* IHPST — CNRS, Université Paris 1 Panthéon-Sorbonne

† SND — CNRS, Sorbonne Université

§ Archives Poincaré — CNRS, Université de Lorraine

Abstract

Scientific models need to be investigated if they are to provide valuable information about the systems they represent. Surprisingly, the epistemological question of what enables this investigation has hardly been investigated. Even authors who consider the inferential role of models as central, like Hughes (1997) or Bueno and Colyvan (2011), content themselves with claiming that models contain *mathematical resources* that provide *inferential power*. We claim that these notions require further analysis and argue that mathematical formalisms contribute to this inferential role. We characterize formalisms, illustrate how they extend our mathematical resources, and highlight how distinct formalisms offer various inferential affordances.

1. Introduction. When analyzing scientific representations, philosophers of science are keen on mentioning that some models provide scientists with "mathematical resources" and "inferential power", but they seldom give a detailed analysis of these notions. This paper is devoted to the discussion of what appears to us as major mathematical resources, namely, formalisms. We thus present an analysis of the notion of formalism as well as examples from which we argue that formalisms should be acknowledged as major units of scientific activity.

We proceed as follows. In Section 2, we briefly review what philosophers of science have to say about mathematical resource and inferential power and observe that it is disappointing. In order to fill the gap we have identified, we put forward in Section 3 the three components we identify within the notion of mathematical resource. Section 4 is devoted to one of these components, namely, formalism. At last, in Section 5, we provide the reader with examples of how the choice of a formalism influences the type of knowledge scientists may draw from their representations.

2. Scientific representations and inferences therefrom. At what conditions can scientific models be used to gain information about target systems? First, a suitable semantic relation between the model and the system(s) that it stands for should obtain, so that by investigating the model, we can make legitimate inferences about its target system(s). This cannot be done unless nontrivial inferences about the model itself, as a mathematical object, can be carried out. Models are usually referred to by proper names (like "Ising model" or "Lotka-Volterra" model") or by expressions that highlight some of their mathematical properties (like "the harmonic oscillator" or "the ideal gas"). There is however more to be learnt about them than their prima facie properties. For example, solving the Ising model reveals more about Ising-like systems than their description as "sets of discrete variables representing magnetic dipole moments of atomic spins that can be in one of two states"; similarly, the mathematical content of an harmonic oscillator goes beyond "being a system that, when displaced from its equilibrium position, experiences a restoring force that is proportional to the displacement". Philosophers of science are aware of the need to investigate the epistemology of models and how we find out about concealed truths about model systems (Frigg,

2010, 257) but are surprisingly silent about how it is actually performed.[1] They are content with saying that the model is "manipulated" (Morgan and Morrison, 1997, chapter 2, passim) or that we can "play" with it (Hughes, 2010, 49), which are suggestive, but metaphoric characterizations.

Surprisingly, even accounts of applied mathematics and scientific representation that give central stage to their inferential role hardly analyze how it is fulfilled and which elements of the models contribute to it. Let us illustrate this point with Bueno's and Colyvan's work. They claim that "the fundamental role of applied mathematics is inferential" (Bueno and Colyvan, 2011, 352) and accordingly propose an "inferential conception" of the application of mathematics that extends Hughes' three-step DDI account of scientific representation (see below).[2] First, a "mapping from the empirical set up to a convenient mathematical structure" (ibidem, 353) is established (immersion step); by doing so, it becomes possible "to obtain inferences that would otherwise be extraordinarily hard (if not impossible) to obtain" (ibidem, 352) (derivation step); finally, the mathematical consequences that were obtained are interpreted step in terms of the initial empirical set up (ibidem, 353) (interpretation step). Bueno and Colyvan further highlight the importance of the inferential role of mathematics for mathematical unification, novel predictions by mathematical reasoning or mathematical explanations (ibidem, 363). However, the analysis of how this inferential role is carried out shines by its absence. Bueno and Colyvan mostly analyze mathematical resources in a semantic perspective[3] and insist on the difference in content and interpretation that these make possible, e.g., when "mathematics provides additional entities to quantify

[1] Frigg, while clearly stating the problem, does not really address it and is content with briefly emphasizing the advantages of his fictional account of model concerning the epistemology of models (Frigg, 2010). As to the epistemological section of Frigg and Hartmann's review article about scientific models, it merely points at experiments, simulations, thought-experiment as ways of investigating models (Frigg and Hartmann, 2017).

[2] Suarez's inferential conception (Suarez, 2004) hardly addresses either the question of how inferences from models are actually carried out. For lack of space, we shall not discuss it here.

[3] Their discussion is mostly directed at the shortcomings of Pincock's "mapping account" of the application of mathematics (Pincock, 2004).

over" (complex numbers), or is "the source of interpretations that are physically meaningful" and provide "novel prediction" about physical systems, like with the case of the interpretation of negative energy solutions to Dirac's equation (ibidem, 366).

In another paper, Bueno suggests that results are derived "by exploring the mathematical resources of the model" in which features of the empirical set up are immersed (Bueno, 2014, 379, see also 387) and that results emerge "as a feature of the mathematics" (ibidem) or by using "the particular mathematical framework" (ibidem, 385). What this inferential power of mathematics should be specifically ascribed to remains unclear. Bueno and Colyvan (2011, 352) just claim that the "embedding *into a mathematical structure* makes it is possible to obtain inferences". They also emphasize how, with the help of appropriate idealizations, "the *mathematical model* [can] directly [yield] the results" (ibidem, 360, our emphasis). But elsewhere in the paper, consequences are said to be drawn "*from the mathematical formalism*, using the mathematical structure obtained in the immersion step" (ibidem, 353, our emphasis).

What are we to make of these various claims? A prima facie plausible answer to this question might be that structures and formalisms are the two sides of a same inferential coin. However, this answer is not satisfactory, since, as is well-known, mathematical structures can be presented in different formalisms, which, as we shall see in Section 4, are associated with different inferential possibilities. Another blind spot in Bueno's and Colyvan's account is that while the derivation step is claimed to be "the *key point* of the application process, where consequences from the mathematical *formalism* are generated'' (ibidem, 353), the question of how inferences are drawn with the help of formalisms is left under-discussed.

We draw from this brief analysis of Bueno's and Colyvan's views that the notions of mathematical resource and inferential power, which are commonly used when discussing applications of mathematics, are often mere labels in need of further investigation. Coming back to the seminal ideas presented by Hughes and extended by Bueno and Colyvan is of little help because Hughes' paper lacks precise answers to the following precise questions: What are exactly mathematical resources? What is their inferential power? In his DDI (Denotation, Demonstration, and Interpretation) account of scientific representation, Hughes claims that scientific representations have an ''internal dynamic'', whose effects we can examine (1997, 332), and "contain *resources* which enable us to demonstrate the results we are interested in". A general notion of resource is appropriate to capture the variety of ways in which demonstrations can be

carried out; however, the claim that the deductive power comes from "the *deductive resources* of mathematics they employ" (ibidem, 332) is too vague and is left unanalyzed.

3. Components of mathematical resources. How are the notions of inferential power and mathematical resources to be analyzed? Are they linked to structures or to symbolic systems and formalisms? In this section, we claim that formalisms are an important component of the notions of inferential power and mathematical resource and should be analyzed in their own right.

Let us begin by briefly presenting what are, according to us, the three main components of the notions of mathematical resource and associated inferential power. First, mathematical structures, *to the extent that they are tractable*, are undoubtedly an important part of the mathematical resources that are used in mathematical modeling. As argued by Cartwright, theories are no "vending machines" that "drop out the sought-for representation" (1999, 247); scientific models are no vending machines either and scientists must make the best of the models that they know to be tractable. Accordingly, the content of models often needs to be adapted by means of idealizations, approximations (Redhead 1980), abstractions, by squeezing representations into the straightjacket of a few elementary models (Cartwright, 1981), or by drawing, from the start, on the pool of existing tractable models (Humphreys, 2004, Barberousse and Imbert, 2014).

Second, mathematical knowledge associated with structures is also to be counted as a distinct mathematical resource, which allows for new inferences when it is available. Let us take the well-known example of Koenigsberg's seven bridges. The impossibility of crossing them once and only once in a single trip can be demonstrated by applying a result from graph theory. Similarly, the explanation of the life-cycle of the Magicicada (Baker 2009, Colyvan 2018) is provided by the application of a number-theoretic property of prime numbers to life-cycles of species.

At last, formal settings or formalisms provide languages in which theories are developed, calculations carried out, and inferences drawn from models. Examples of formalisms are Hamiltonian formalism, path integrals, Fourier representation, cellular automata, etc. We provide a detailed analysis of some of these below. Contrary to mathematical structures, formalisms are partly content neutral (though form and content are often intertwined in scientific representations). As providing a partially stan-

dardized way of making inferences, they are important tools for scientists, which in turn justifies considering them as important units of analysis in the philosophy of science. Other authors have started exploring the idea that format matters in scientific activities. Humphreys gives general arguments to this effect and emphasizes the difference between formats that are appropriate for human-made and format that suit computational inferences (2004). Vorms (2009) also emphasizes the general importance of formats of representation when toying with theories or models. Formalisms are a specifically mathematical type of format whose role needs further investigation. This is what we do in the next section.

4. What are formalisms? As briefly stated above, formalisms are mathematical languages that allow one to present mathematical statements or objects and draw inferences about them by means of general inference rules. For example, *Hamiltonian formalism* is one of the formalisms through which scientists may find out means to solve differential equations. *Path integrals* is another formalism of this kind, with the help of which one may also solve (partial) differential equations. Let us illustrate the latter point further: the integral solution of the Schrödinger equation requires using a mathematical object, the *propagator,* whose calculation the path integrals formalism makes easier. *Fourier representation or formalism* enables one to represent mathematical functions as the continuous sum of sine functions (or complex exponential functions), so that harmonic analysis, i.e. the decomposition of a signal in its harmonic frequencies, may be performed. It also provides modelers with a way to express the solutions of some partial differential equations, such as the heat equation. Finally, formalisms like *numerical integrators, cellular automata, lattice Boltzmann methods,* and *discrete variational integrators*, are indispensable in current computational science.

Formalisms consist in the following elements:

i. elementary symbols;

ii. syntax rules that determine the set of well-formed expressions;

iii. inference rules;

iv. a partly detachable interpretation, both mathematical and physical.

Their use is facilitated by

v. translation rules that indicate how to shift from one formalism to another.

Let us illustrate these elements by discussing in more detail the above examples. In the Hamiltonian formalism, elementary symbols are used for a variable and its conju-

gate momentum: "(q, p)", or for Poisson brackets ''{.,.}''. Among the syntax rules that are specific to Hamiltonian formalism, some allow one to rewrite Hamilton equations by using the canonical variables. Inferences rules allow the users to use action-angle variables (I, *theta*) and to solve equations by using these coordinates because this change of variables opens the possibility to deal with integrable systems, thus providing a systematic method to solve *exactly*, i.e., in closed forms, differential systems like the simple pendulum, and more generally, any 1D-conservative system. Indeed, due to this change of variables, one takes full advantage of the existence of conserved quantities in mechanical systems, which are then used as variables (actions) in Hamilton equations. This allows constructing the solution of the equations by "quadrature'' (Babelon et al. 2003, chapter 2). An example of a translation rule is the Legendre transform that allows one to shift to Lagrangian formalism. Similarly, in the case of Fourier transforms, an elementary specific symbol is $f^\wedge$, which corresponds to the Fourier transform of the function *f*. Scientists use sets of rules that describe the Fourier transforms of some typical functions, such as the constant function, the unit step function, and the sinusoids, but also rules for the convolution product, viz. the Fourier transform of the convolution $f \circ g$ is the product of Fourier transforms of *f* and *g*: $(f \circ g)^\wedge = f^\wedge . g^\wedge$, so that solutions of equations may be found within Fourier space. An inverse Fourier transform is also defined, which enables one to move back from the Fourier transform $f^\wedge$ to the function *f* (this is again a translation rule).

As emphasized above, formalisms are (partly) content neutral and thus "exportable", even though they usually come with a privileged physical interpretation. As a matter of fact, most formalisms have been developed within a peculiar modeling context or are linked to a physical theory. From this origin, the most successful ones may become autonomous and depart from their original, physical interpretation. For example, Hamiltonian formalism was initially developed in the context of classical mechanics but is nowadays autonomous and used in other physical contexts. Path integrals originally come from the study of Brownian motion (Wiener 1923) and quantum mechanics (Feymann 1942) but are currently used in other fields like field theory and financial modeling.

The mathematical interpretation of formalisms may sometimes be detachable. For example, the transition rules associated with cellular automata (see below) do not have any obvious mathematical interpretation. Further, although some formalisms are linked to acknowledged mathematical theories (e.g., the Fourier formalism is linked to

the theory of complex functions), they differ from genuine mathematical theories, as shown by the example of path integrals, in which the formalism is used in the absence of any uncontroversial mathematical theory that could back it up. The definition of a path integral:

$$K(b,a) = \int_a^b e^{\frac{2i\pi}{h}\int_{t_b}^{t_a} L\,dt}\, Dx(t)$$

requires using a measure "Dx", to which no general, rigorous definition can be given yet. This mathematical concern does not prevent physicists from using path integrals anyway, as testified by the following quote: "The question of how the path integral is to be understood in full generality remains open. Given this, one might expect to see the physicists expending great energy trying to clarify the precise mathematical meaning of the path integral. Curiously, we again find that this is not the case" (Davey 2003, 450).

Let us finally emphasize that formalisms also differ from formulations of physical theories and allow philosophers of science to address different philosophical problems. Formulations of theories, in particular axiomatic ones, are explored when questions about conceptual content and metaphysical implications are raised. They pertain to foundational issues. Whether a given formulation involves calculus is a peripheral issue in this context. By contrast, the primary virtue of a formalism is to allow modelers to draw actual inferences from a theory or model. The inferential rules it contains are more important than the mathematical rigor of the language in which it is expressed.

5. Choosing a formalism. So far, we have argued that the inferential power that is required to explore models is partly brought about by formalisms, and we have given examples thereof. Accordingly, formalisms have to be carefully examined by philosophers of science if they are to provide a fine-grained analysis of how scientific knowledge is produced in practice. We now aim to show that there is no unique description of formalism-rooted inferential power since different formalisms allow for different types of inferences and are adapted to different types of inquiries. We do so by providing examples of these differences and of the factors that guide scientists when choosing the formalism that is best suited to the task at hand.

How do scientists decide which formalism to use in a given inquiry? The choice may first depend on the type of models at hand. For example, the path integral formalism is

well adapted to solve systems with many degrees of freedom (Zinn-Justin 2009) and makes ''certain numerical calculations in quantum mechanics more tractable'' (Davey 2003, 449). Lagrangian formalism offers a well-suited framework to solve equations describing constrained systems (Goldstein 2002, 13, Vorms 2009, 15). Fourier representation allows one to solve, e.g., the differential equations describing the time evolution of electrical quantities in networks. In this case, differential equations are transformed into *algebraic equations* on variables in Fourier space, which may be easier to solve. Finally, with the change of action-angle variables, Hamiltonian formalism potentially provides exact solutions for integrable systems, which have as many independent conserved quantities as degrees of freedom.

The use of a particular formalism is also guided by epistemic goals. Depending on the chosen formalism, different kinds of properties, general (e.g. periodicity, symmetry) or particular (dynamical), may be inferred from the same model. Let us illustrate this point with the example of prey-predator models in ecology. Among these, some obey Lotka-Volterra (LV) equations and represent transforming populations with a system of two coupled equations. If they are investigated within the Hamilton formalism, *general properties* of these models can be found without setting initial conditions or numerical values for the involved parameters. The reframed models can indeed be shown to be integrable, like the simple pendulum in classical mechanics. Dutt explicitly emphasizes the advantages of using this formalism for a two-species LV system:

> ''In dealing with the problems involving *periodicity*, the Hamilton-Jacobi canonical theory has a distinct advantage over the conventional methods of classical mechanics. In this approach, one introduces action and angle variables through canonical transformations in such a way that the angle variable becomes cyclic. One then obtains the frequency of oscillation by taking the derivative of the Hamiltonian with respect to the action variable. One may thus *bypass the difficulty* in obtaining the complete solutions of the equations of motion, *if these are not required*.'' (Dutt, 1976, 460, our emphasis)

LV models can also be solved with the help of computers and generic numerical integrators when the aim is to obtain particular dynamics for specific values of parameters and initial conditions. Such numerical solutions of the LV model can also be provided by specific formalisms, such as discrete variational integrators (Krauss 2017, 34; Tyranowski 2014, 149). In that case, discrete equations are derived from a discrete least action principle, which is well-suited to conservative systems, like the LV sys-

tem. Discrete variational integrators allow for the preservation of general properties like the conservation of global quantities, viz. energy, momenta, and symplecticity. This discrete formalism comes with mathematical constraints on the discretization of time since the time step has to be adaptive in order to guarantee the conservation of global quantities (Marsden & West 2001, Section 4.1).

Finally, let us mention that LV models can also be studied by using *cellular automata* (CA) and associated formalism, with the following advantages:

> [a rather general predator-prey model] is formulated in terms of automata networks, which describe more correctly the *local character* of predation than differential equations. An automata network is a graph with a discrete variable at each vertex which evolves in discrete time steps according to a definite rule involving the values of neighboring vertex variables. (Ermentrout and Edemstein-Keshet 1993, 106)

On the one hand, CA are discrete dynamical systems, but on the other, they are also a nice means to practice science with the help of a computationally simple formalism (in terms of transition rules). They can be extremely powerful. For example, rule 110 is Turing complete and, like lambda-calculus, can emulate any Turing machine and therefore complete any computation. In contrast with the case of Hamilton formalism, CA-based inferences from prey-predator models are carried out for specific values and parameters. As CA are described by local rules, these inferences merely pertain to local variations in the model. However, the simplicity of these rules is a tremendous advantage for modeling and code-writing. For instance, CA allow one to easily add rules for the pursuit and evasion of populations as well as rules for age variation (Boccara et al. 1993, Ermentrout and Edemstein-Keshet 1993, see also Barberousse and Imbert 2013 for an analysis of CA as used in fluid dynamics and compared with Navier-Stokes based methods).

Let us now turn to a different example illustrating how different the epistemological effects of using this or that formalism may be. Crystals are currently modeled as lattices that come under two forms, *lattices in real space* and *lattices in reciprocal space*. Each is associated with a specific formalism. Within the *real space lattice* formalism, crystals are described with a vector R expanded on a vector basis (a_1, a_2, a_3) which corresponds to crystal directions, and *alpha*, *beta*, *gamma* are the corresponding angles. Inferences about *symmetry* of crystals are usually made within this type of representation since the real space is well adapted to studying discrete translations and rotations.

Crystals can also be described with the help of a vector R* in a *lattice in reciprocal space*. There is a clear correspondence between the two spaces since they are dual. Given R in the real space, we can derive R* in the reciprocal space, and conversely. The two spaces are related by a Fourier transform. However, the *reciprocal space* can be more convenient because inferences about *diffraction and interference patterns* are easier to carry out in the Fourier representation. As stressed by Hammond in a textbook of crystallography:

> the reciprocal lattice is the basis upon which the geometry of X-ray and electron diffraction patterns can be most easily *understood* and […] the electron diffraction patterns observed in the electron microscope, or the X-ray diffraction patterns recorded with a precession camera, are simply sections through the reciprocal lattice of a crystal (Hammond 2009, 165).

This example shows that facilitating inferences may have various epistemological effects. Some are relevant to computational aspects and the predictions or explanations that scientists are able to produce in practice. Others pertain to the way scientists understand and reason about models and their target systems. This example also shows how different epistemic goals (symmetry-oriented vs. interference-oriented investigations of crystals) determine which formalism is chosen.

Overall, the above shows that formalisms not only have an important impact on the amount of results scientists may produce, but also on the types of results that are attainable. The examples we have discussed also highlight that the existence of a variety of formalisms is a source of epistemic richness and enhanced inferential power for scientists because it provides them with multiple ways of investigating the same mathematical structures or structures that are related by suitable morphisms.

6. Conclusion. The above proposals are meant to contribute to the epistemological question of what provides models with inferential power and helps scientists succeeding in their inquiries. We have shown that some of this inferential power is brought about by the formal symbolic tools that scientists use to present and investigate mathematical models. Our second claim is that all formal settings do not enable the same types of inferences nor are suited to all epistemic goals. Accordingly, a fine-grained analysis of the conditions of scientific progress needs, among other things, to focus on formalisms.

Our epistemological analysis is not tied to any particular theory of scientific representation. However, by showing that inferences actually hinge on choice of formalism, it suggests that a theory of scientific representation that is cashed out in terms of structures is too abstract to account for the various ways equations are solved in practice and information extracted from scientific models.

References

Babelon Olivier, Bernard Denis, and Talon Michel. 2003. *Introduction to classical integrable systems*, Cambridge: Cambridge University Press.

Baker, A. 2009. "Mathematical Explanation in Science". *British Journal for the Philosophy of Science* 60 (3): 611–633.

Barberousse, Anouk, and Cyrille Imbert. "New Mathematics for Old Physics: The Case of Lattice Fluids." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 44 (3) : 231–41.

Barberousse, Anouk, and Cyrille Imbert. 2014. "Recurring Models and Sensitivity to Computational Constraints" *The Monist* 97 (3): 259–79.

Boccara Nino, Roblin O. and Roger Morgan. 1994. Automata network predator-prey model with pursuit and evasion, *Physical Review E* 50 (6): 4531–41

Bueno, Otávio, and Mark Colyvan. 2011. "An Inferential Conception of the Application of Mathematics". *Noûs* 45 (2): 345–74.

Bueno, Otávio. 2014. "Computer Simulations: An Inferential Conception". *The Monist* 97 (3): 378–98.

Cartwright, Nancy (1983). *How the Laws of Physics Lie*. Clarendon Press, Oxford.

Cartwright, Nancy.1999. "Models and the Limits of Theory: Quantum Hamiltonians and the BCS Models of Superconductivity". In *Models as Mediators*, ed. Mary S. Morgan and Margaret Morrison Morgan, Cambridge: CU Press: 241–81.

Colyvan, Mark. Forthcoming. "The Ins and Outs of Mathematical Explanation", *Mathematical Intelligencer*.

Davey Kevin. 2003. "Is Mathematical Rigor Necessary in Physics?" *The British Society for the Philosophy of Science*, 54(3): 439–463

Dutt Ranabir. 1976. "Application of the Hamiltonian-Jacobi Theory to Lotka-Volterra Oscillator", *Bulletin of Mathematical Biology*, 38: 459–465.

Ermentrout G. Bard and Edemstein-Keshet, Leah. 1993. "Cellular Automata Approaches to Biological Modeling". *Journal of Theoretical Biology* 160: 97–133.

Feynman, Richard. P. 1942. "The Principle of least action in quantum mechanics", *PhD. diss.*, Princeton University.

Frigg, Roman. 2010. "Models and Fiction". *Synthese* 172 (2): 251–68.

Frigg, Roman, and Stephan Hartmann. 2017. "Models in Science." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/models-science/.

Goldstein, Herbert. 2002. *Classical Mechanics*. Reading, Mass: Addison-Wesley.

Hammond, Christopher. 2009. *The Basics of Crystallography and Diffraction*, Oxford University Press.

Hughes, Robert I.G. 1997. "Models and Representation". *Philosophy of Science* (Proceedings): 64: S325–S336.

Hughes, Robert I.G. 2010. *The Theoretical Practices of Physics: Philosophical Essays*. Oxford: Oxford University Press.

Humphreys, Paul. 2004. *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Kraus, Michael. 2017. "Projected Variational Integrators for Degenerate Lagrangian Systems", preprint: https://arxiv.org/pdf/1708.07356.pdf

Marsden Jerrold E. and West Matthew. 2001. "Discrete Mechanics and Variational Integrators", *Acta Numerica*, 10: 357–514.

Morgan, M., and Margaret Morrison (1999). *Models as Mediators*. Cambridge University Press.

Pincock, Christopher. 2004. "A new perspective on the problem of applying mathematics", *Philosophia Mathematica* 3 (12), 135-61.

Redhead, M. 1980. "Models in Physics", *The British Journal for the Philosophy of Science*, 31(2): 145-163

Suarez, Mauricio. 2002. "An Inferential Conception of Scientific Representation", *Philosophy of Science* 71 (5): 767-779

Tyranowski Tomasz. M. 2014. "Geometric integration applied to moving mesh methods and degenerate Lagrangians". Ph.D. diss., California Institute of Technology.

Vorms, Marion. 2011. "Formats of Representation in Scientific Theorizing." In *Models, Simulations, and Representations*, edited Paul Humphreys and Cyrille Imbert. Routledge.

Wiener, Norbert. 1923. "Differential space". *Journal of Mathematical Physics* 2: 131-174.

Zinn-Justin Jean. (2009), Path Integral, *Scholarpedia*, 4(2): 8674.

Representation Re-construed: Answering the Job Description Challenge with a Construal-based Notion of Natural Representation

Abstract: Many philosophers worry that cognitive scientists apply the concept REPRESENTATION too liberally. For example, William Ramsey argues that scientists often ascribe natural representations according to the "receptor notion," a causal account with absurd consequences. I rehabilitate the receptor notion by augmenting it with a background condition: that natural representations are ascribed only to systems construed as organisms. This Organism-Receptor account rationalizes our existing conceptual practice, including the fact that scientists in fact reject Ramsey's absurd consequences. The Organism-Receptor account raises some worrying questions, but as a more faithful characterization of scientific practice it is a better guide to conceptual reform.

Abstract: 100 words

Total: 4,995 words

1. Introduction. There is a common complaint among philosophers that scientists use the word "representation" too liberally. Representation is often contrasted with indication: representation is a distinction achieved by maps, linguistic performances, and thoughts, whereas indication is a less-demanding state achieved by thermostats, which indicate ambient temperature, and refrigerator lights, which indicate whether the door is open (Dretske 1981; Cummins and Poirier 2004). However, cognitive scientists often ascribe representations when it seems that mere indication is all that is called for. We commonly say that hidden layers in a neural network represent concepts, or that neurons in V1 represent visual edges, because they reliably respond differently to the circumstances they are said to represent (Ramsey 2007, 119–20; cf. Hubel and Wiesel 1962). But these "representations" are thin-blooded compared to paradigmatic conventional representations. For example, they cannot be invoked in the absence of an appropriate stimulus. So are cognitive scientists conceptually confused? Do they exaggerate their claims? And if the natural representations posited by cognitive scientists aren't genuine representations, is the cognitive revolution dead?

William Ramsey provides an excellent book-length exploration of these worries, articulating a qualified pessimism about their answers:

> …we have accounts that are characterized as "representational," but where the structures and states called representations are actually doing something else. This has led to some important misconceptions about the status of representationalism, the nature of cognitive science and the direction in which it is headed. (2007, 3)

Ramsey describes the "job description challenge": to give an account of the distinctive properties of representations in virtue of which appealing to them serves a special

explanatory role. If the job description challenge can be met, then we can formulate a plan for conceptual reform.

I undertake Ramsey's challenge, but with a metadiscursive twist: I describe the Organism-Receptor account, which articulates conditions for ascribing representations, in virtue of which such ascriptions achieve a special explanatory purpose. The account is merely suggestive about the properties that distinguish first-order representational states from non-representational states; it says more about the mental state of the ascriber than about the representation-bearing system. However, the Organism-Receptor account provides a more adequate characterization of scientists' practice than Ramsey's.

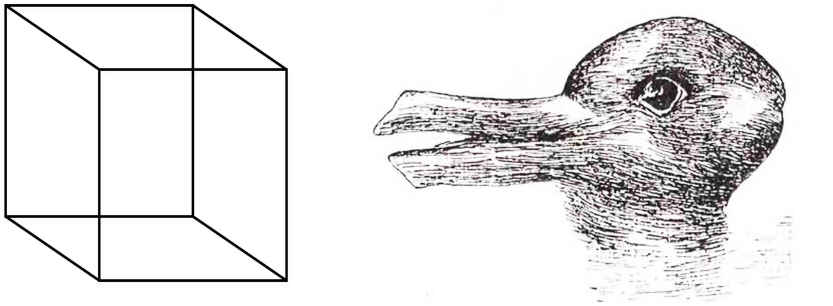
My main aim in this paper is to push back against pessimistic evaluations of the existing practice of representation-ascription in cognitive science, like Ramsey's. I will focus on Ramsey's critique of the "receptor notion," a flawed causal theory of representation that he attributes to some cognitive scientists. Ramsey argues that the receptor notion has absurd consequences, although scientists do not accept them. By augmenting the receptor notion with a construal-based background condition, I can explain why scientists do not draw these absurd conclusions. Whereas Ramsey's pessimistic account of scientists' practice of ascribing representations finds it wanting and is extensionally inadequate, mine rationalizes our extant conceptual practice (though that practice is not beyond criticism). I conclude that my apologetic account is a more charitable and adequate interpretation of existing scientific practice than Ramsey's.

2. Ramsey on the "Receptor Notion." Ramsey argues that natural representations in cognitive science are often ascribed according to the "receptor notion," a crude causal theory of representation. According to the receptor notion, a state *s* represents a state of affairs *p* if *s* is regularly and reliably caused by *p* (2007, 119).

Ramsey claims that the receptor notion is what justifies the ascription of representations to cells in V1 that detect visual edges, cells in frog cortex that detect flies, and the mechanisms in Venus flytraps that cause their "jaws" to close (119–23). Ramsey argues that this receptor notion is too liberal to be useful to scientists. For example, it is susceptible to the "disjunction problem" (Fodor 1987): since frog neurons respond reliably to visual stimulation by flies *or* (say) BBs, we should say that the content of the representation is *fly-or-BB*, rather than *fly*. Likewise, Venus flytraps represent objects in a particular range of sizes rather than *edible insects*, and the human concept GOAT represents *goats-or-weird-looking-sheep*. Such disjunctive content-ascriptions are usually considered absurd. Absent a clever fix, we must embrace unwieldy, disjunctive contents for representations or we must reject the receptor notion (Ramsey, 129).

Dretske's (1988) teleofunctional theory of representation is a sophisticated twist on the receptor notion that avoids the disjunction problem. On Dretske's view, a representational state must not only be causally dependent on the state of affairs it represents, but must serve a function for its containing system in virtue of this causal dependency. This extra condition motivates constraints on representational content that eliminate problematic disjunctive contents. Dretske's theory is subject to some subtle criticisms that I will discuss in Section 6, but the Organism-Receptor account will preserve some of the teleological character of Dretske's theory.

Ramsey's most compelling objection to the receptor account, including Dretske's sophisticated version, is that it justifies ascribing representational contents to states that are not, in fact, representational: smoke "represents" fire since the latter causes the former. Likewise, the firing pin of a gun "represents" whether the trigger is depressed, and rusting iron "represents" the presence of water and oxygen (138–47). Ramsey claims, plausibly, that these are absurd consequences. I find Ramsey's reductio



Ambiguous figures. Left: The Necker cube. Right: The duck-rabbit (image from Jastrow 1899).

compelling, but reject a different premise than he does. Rather than conclude that cognitive scientists have a bad conceptual practice, I question whether his characterization of the receptor notion is a charitable understanding of what happens in cognitive science. After all, cognitive scientists do not generally claim that GOAT denotes *goats-or-sheep* (at least for competent judges of goathood), or that firing pins represent anything.

3. A Construal-based Notion of an Organism. I argue that something like the receptor notion can be salvaged if being a receptor is contextualized in terms of construal. Construal (also called "seeing-as") is a judgment-like attitude whose semantic value can vary licitly independently of the state of affairs it describes. For example, we can construe an ambiguous figure like the Necker cube as if it were viewed from above or below, or the duck-rabbit as if it were an image of a duck or of a rabbit (Roberts 1988; see also Wittgenstein 1953). We can construe an action like

skydiving as brave or foolhardy, depending on which features of skydiving we attend to.

On a construal-based account of conceptual norms, a concept (e.g. REPRESENTATION) is ascribed relative to a construal of a situation. For example, perhaps I fear something only if I construe it as dangerous to me or detrimental to my ends (Roberts 1988). Daniel Dennett's (1987) intentional stance is a more familiar example: according Dennett, a system has mental states if and only if we construe it in such a way that its behavior is explainable in terms of a belief-desire schema.

I propose that construing something as an organism involves construing it such that it has goals and behavior, and believing that it has mechanisms that promote those goals by producing that behavior. More precisely:

> **Organism-Construal.** A subject *a* construes a system *x* as an organism in a context[1] *c* if and only if, in *c*,
>
> (O1) *a* attributes a set of goals *G* to *x*,
>
> (O2) *a* attributes a set of behaviors *B* to x,
>
> (O3) *a* believes that the elements of *B* function to promote elements of *G*,
>
> (O4) *a* believes that *x* possesses a set of mechanisms *M*, and
>
> (O5) *a* believes that the elements of *M* collectively produce the elements of *B*.

My main argument does not rely on all the details of Organism-Construal; it could be replaced by a different explication of what it is to see something as an organism. But Organism-Construal captures an intuitive notion of a critter. First of all, we normally take living critters to have goals, such as survival and reproduction, and behaviors that

[1] The relevant notion of a context is something like MacFarlane's (2014) "context of assessment."

promote those goals. However, Organism-Construal does not require that an organism really have goals (whatever that involves) or exhibit behavior (however that's distinguished from other performances). To see something as an organism according to Organism-Construal, the construing subject need only *attribute* goals to the system, and see some of its performances as behaviors that promote those goals. Such goals could include relatively specific aims such as locating food, getting out of the rain, or driving home. We sometimes also attribute goals and behaviors to non-living things, such as automated machines. For example, we might say that a robot vacuum has the goal of cleaning the floor, which it accomplishes by sucking up dust. Or I might say that my GPS navigation computer is trying to kill me, which it accomplishes by consistently giving me directions that lead me through strange, dangerous backroads. Condition (O3) is expressed in terms of belief instead of attribution, meaning that the construing subject must sincerely believe that an organism's putative behaviors function to promote its putative goals. When and insofar as someone construes a system in this way, the conditions (O1)–(O3) above are satisfied.

Conditions (O4)–(O5) require that the system's behavior be explainable by appeal to mechanisms. "Mechanisms" here should be understood in roughly the sense meant by the new mechanists (Machamer, Darden, and Craver 2000; Bechtel and Abrahamsen 2005; Craver 2007): organized structures of component parts and operations that produce a phenomenon, and the description of which is an explanatory aim of some scientific projects. Much explanation in biology and neuroscience plausibly follows a mechanistic model, and likewise in cognitive science. Daniel Weiskopf (2011) has argued that cognitive explanations are not properly mechanistic, but even on his view cognitive explanations are extremely similar to mechanistic ones, distinguishable only because the relationship between components of cognitive models and their physiological realizers is relatively opaque. Regardless, cognitive scientists use the word "mechanism" to refer to the referents of their models,

just as biologists and neuroscientists do. I am more moved by the similarities between the biological and the cognitive sciences than the differences. Therefore, like Catherine Stinson (2016), I acknowledge Weiskopf's concerns but nevertheless adopt the language of "mechanisms."

Not all of a system's mechanisms function to produce behavior. For example, biological organisms have metabolic and other mechanisms that maintain bodily integrity. Such mechanisms may need to function correctly as a background condition for the organism to behave, but scientists do not typically take behavioral patterns to be the explanandum phenomena of such mechanisms. Let us call mechanisms that do contribute to the explanation of behavior *behavioral mechanisms*. As for what it means for a system to "possess" a mechanism, a mereological criterion will do for now: the mechanism must be a part of the system. Condition (O5) is meant to limit the mechanisms in the set *M* to behavioral mechanisms.

So far so abstract; let's consider an example. The robot Herbert was designed to wander autonomously through the MIT robotics lab, avoiding obstacles, and collecting soda cans with its arm (Brooks, Connell, and Ning 1988). Herbert can be construed as an organism, even though it is not alive, as long as one (O1) attributes goals, like avoiding collisions and collecting soda cans, to Herbert, (O2) sees some of Herbert's performances as behaviors, (O3) believes that Herbert's behaviors promote its goals, and (O4) believes that Herbert possesses mechanisms that (O5) explain its behavior. Herbert does possess mechanisms for accomplishing goals; it is equipped with sensors, computers, and motors that coordinate its locomotion and its grasping arm. And most people readily anthropomorphize Herbert enough to see it as a goal-directed, behaving system (pace Adams and Garrison [2013], who insist that Herbert has its designers' goals, but no goals of its own). Anyone willing to engage in the imaginative attribution of goals and behavior to Herbert can see Herbert as an organism, even if on reflection they believe Herbert is not literally an organism. The

willingness to ascribe representations to a system plausibly waxes and wanes along with one's willingness to construe the system as an organism in something like the sense described above. There are psychological limits on the willingness to attribute goals and behaviors to systems relatively unlike animals, and these limits may vary between individuals.

4. The Receptor Notion Re-construed. Returning now to the receptor notion of natural representation, I suggest that it can be augmented in the following way:

> **Organism-Receptor.** A state *s* represents a state of affairs *p* if
>
> (R1) *s* is regularly and reliably caused by *p*, and
>
> (R2) *s* is a functional state of a behavioral mechanism possessed by an organism.

Organism-Receptor is not a construal-based explication, but it depends on a construal-based account of ORGANISM. It preserves the spirit of Ramsey's receptor notion, with the added condition that representations be ascribed to parts of systems construed as organisms. Representation-ascriptions guided by Organism-Receptor inherit their plausibility from the plausibility of the corresponding construal of some system as an organism. Most accounts of cognitive representation require there to be a representational subject of some kind (e.g. Adams and Aizawa 2001; Rupert 2009; Rowlands 2010), and on Organism-Receptor the organism serves this role. We can constrain the acceptable contents of these representations by requiring they correspond to descriptions of *p* according to which *p* is relevant to the pursuit of an organism's goals. This appeal to goals is not ad hoc, since according to Organism-Receptor representations are ascribed to organisms, i.e. systems to which we've already attributed a set of goals. Thus, like Dretske's (1988) and Millikan's (1984)

teleofunctional accounts, this construal-based account addresses the disjunction problem by appealing to goals of organisms.

The metadiscursive job-description challenge is to provide criteria of ascription for representations, in virtue of which representation-ascriptions achieve some explanatory purpose. I have provided criteria of ascription, so what is their purpose? On Donald Davidson's (1963, 5) account of intentional action, actions are performed under the guise of a privileged description (or set of descriptions). Davidson flips the light switch in order to turn on the light, but not in order to alert the prowler outside (whose presence is unknown to Davidson) that he is home, though he also does the latter. Davidson calls this feature of action its "quasi-intensional character." Behavioral mechanisms also have something like a quasi-intensional character, since there are privileged descriptions that make explicit how they and their components contribute to an organism's capacity to pursue its goals. For example, edge-detecting cells in V1 fire in order to identify boundaries in an organism's environment, not to consume glucose, though they also do the latter. The use of representation-talk by cognitive scientists, as licensed by Organism-Receptor, is a way to habitually mark these privileged descriptions and distinguish them from other descriptions of the same states or events. And since cognitive science is concerned with the functional structure of behavior-coordinating mechanisms rather than other features of cognitive systems, it is easy to see why representation—even in this relatively thin sense—has always been the dominant theoretical perspective in cognitive science. This focus on quasi-intensional characterization may even be what makes the cognitive scientific perspective distinctive (on scientific perspectives, see e.g. Giere 2006).

The Organism-Receptor account provides us with resources to salvage the receptor notion from Ramsey's reductio. It is plausible to suppose that cognitive scientists generally ascribe natural representations to systems against an imaginative

background like this. After all, most cognitive science concerns the mechanisms of living systems, especially animals (except in computer science and some computational modeling, where the object of attention is a formal object like a connectionist network that is presumed to be analogous in some way to such a mechanism). Such systems are easily construed as organisms in the sense of Organism-Construal. Non-living things and even non-animals are in general more difficult to construe as organisms in that sense, since they are often perceived to lack goals, the capacity to behave, or both.

5. The Organism-Receptor Notion in Context. Consider a strong case of representation, like fly-detecting cells in frog visual cortex. We construe frogs as systems that exhibit goal-directed behavior and believe they possess mechanisms that explain that behavior. Frog visual cortex contains mechanisms that (along with other mechanisms) explain behaviors like fly-catching. When we identify cells in frog visual cortex that fire in response to the visual presence of flies (or fly-like objects), we ascribe representational properties to those cells. The contents we ascribe to representations in frog visual cortex are constrained by the goals we attribute to frogs. *That a small insect is present* is a suitable content because flies can be consumed for energy; *that a wiggly BB is present* does not have this significance for frogs, although BBs may be indistinguishable from insects by the mechanisms in the frog's visual cortex. Nevertheless, the relationship between fly-presence and the frog's goals provide a ground for privileging non-disjunctive descriptions of representational content.

The Organism-Receptor account also explains why liminal cases of representation, like the case of Herbert, are liminal. We can say that Herbert represents such states of affairs as the presence of obstacles and soda cans, because states of Herbert's sensors are regularly and reliably caused by those states of affairs.

And we can ascribe contents to representations by drawing on descriptions of Herbert's environment that relate to the goals we ascribe to Herbert. However, our willingness to take these representations seriously as natural representations that bear content intrinsically covaries with our willingness to take Herbert seriously as an organism. We are not as comfortable attributing genuine goals and behaviors to Herbert as we are attributing goals and behaviors to frogs.[2]

Finally, absurd cases like the firing pin can be excluded (for the most part) since guns are not easily construed as "organisms." Firearms are difficult to anthropomorphize, since they do not exhibit autonomous behavioral dynamics and we don't normally see them as having goals of their own. It is not *impossible* to ascribe goals to weapons or other tools, but the ascription of folk-psychological properties to tools, like the folk ascription of a bloodthirsty disposition to a sword, generally depends on the way a tool influences its users' behavior. (I suspect this dependence might offer some novel explanations of why Clark and Chalmers' [1998] extended cognition hypothesis is attractive to some.) The attribution of autonomous behaviors to tools like swords is fanciful. Perhaps we might imagine a tool exhibits psychic "behavior," but anyway we do not believe that swords possess mechanisms that produce this "behavior" (though if we did, such a construal would be more compelling). If the firing pin of a gun is not a component of a behavioral mechanism, it cannot represent anything according to the Organism-Receptor account.

So the Organism-Receptor account licenses an ascriptive practice that resembles the crude receptor notion when the role of construals is not made explicit. It is unusual in that it inverts Ramsey's preferred order of ascription: Ramsey wishes to

---

[2] Notably, Rodney Brooks himself does not claim that it is proper to ascribe representational capacities to Herbert (Brooks, Connell, and Ning 1988; Brooks 1991), but Brooks plausibly had in mind a more demanding account of representation.

ascribe cognitive structure to systems in virtue of their representational structure (see e.g. Ramsey, 222–235), whereas I suggest that we in fact ascribe representational structure in virtue of seeing a system as a system with goal-directed behavior, i.e. as a potentially cognitive system.

6. Worries. Since the Organism-Receptor account shares a certain teleological character with Dretske's account, I will discuss Ramsey's two most developed objections to Dretske, along with other worries specific to the Organism-Receptor account. First, Ramsey objects that Dretske's account is question-begging with regard to the job-description challenge. Roughly, teleological normativity (i.e. functioning and malfunctioning) is not sufficient to explain intentional normativity (i.e. representation and misrepresentation), and since Dretske provides no satisfying criteria for what it is for a state to function as a representation, he cannot bridge that gap (Ramsey 2007, 131–2). But the Organism-Receptor account has more resources than Dretske's teleofunctionalism. Construing a system as an organism involves construing it as exhibiting behavior, which allows us to distinguish behavioral mechanisms from other mechanisms. On the Organism-Receptor account, misrepresentations are malfunctions of behavioral mechanisms (like frog vision), but not of other mechanisms (like a frog's circulatory system or a gun's firing mechanism).

My reply invites a rejoinder: on the Organism-Receptor account the functional roles of representations will be extremely diverse, and representations will be common. They will not just include IO-representation and S-representation (roughly, information-processing relata and models for surrogative reasoning; Ramsey 2007, 68ff.), which Ramsey and most cognitive scientists regard as genuinely representational. They will also include more controversial varieties of "representation," such as Millikan's (1995) "pushmi-pullyu" representations: Janus-faced mechanistic components that simultaneously indicate a state of affairs and cause

an adaptive or designed response. In other words, representations will include what Ramsey calls "causal relays" like the firing pin in a gun, the inclusion of which in the extension of REPRESENTATION was the ground for his reductio! However, the absurd cases can be avoided. The firing pin case is excluded because guns are poor examples of organisms. And pushmi-pullyu representations include cases with significant intuitive appeal to many scientists, like the predator calls of vervet monkeys (Millikan 1995; cf. Seyfarth, Cheney, and Marler 1980). While this conception of representation has a more liberal extension than Ramsey is comfortable with, it is liberal enough to explain common representation-ascriptions in cognitive science without being so liberal as to countenance absurd cases like Ramsey's firing pin, so I submit it is adequate to scientific practice.

Ramsey's second objection is that Dretske is committed to a false principle: that if a component is incorporated into a mechanism because it carries information, then its function is to carry information (132–9). However, the Organism-Receptor account constrains the causal dependence criterion (R1) by relying on construals of systems as organisms instead of teleofunctional commitments. The account I describe is not committed to Dretske's principle, and therefore is not subject to this objection.[3]

Nevertheless, one might worry whether the organism criterion (R2) is a suitable condition on representation-ascription. I suggested five conditions (O1)–(O5) on what can be seen as an organism, but conditions (O1) and (O2) are fairly unconstrained. There are psychological limitations on when goals or behaviors can be plausibly attributed to a system, but what are those limits? And what factors influence interpersonal variability in willingness to make these attributions? The reason this practice isn't bonkers is that it coheres with the explanatory purpose of

[3] Ramsey's discussion is rich and worthy of deeper engagement than this, but for reasons of space I leave the matter here.

representation-ascriptions: to make explicit the quasi-intentional character of behavioral mechanisms. Nevertheless, we should hope that these psychological limitations are vindicated by more principled considerations. Criticism is warranted if scientists attribute goals and behaviors when they should not. There is some extant work on the proper norms ascribing goals to organisms (e.g. Shea 2013; Piccinini 2015, chap. 6), but little serious work on how to understand the concept of BEHAVIOR in the context of cognitive science. We should worry about the practice of ascribing natural representations if scientists construe things that are not cognitive systems as "organisms." Indeed, we might indeed worry that many cognitive scientists misuse the concept COGNITION, given the intense disagreements over its extension (see e.g. Akagi 2017). However, my present aim is not to evaluate scientific practice, but to describe it faithfully (with the hope that a more satisfactory evaluation will follow).

Another worry about construal-based accounts is that they entail an unattractive anti-realism: if representations and their contents only exist relative to construals, they are mind-dependent rather than objective, right? This worry is unfounded. I am undertaking a modified version of Ramsey's job description challenge: my aim is to describe the ascription of representations in virtue of which they serve an explanatory purpose, not to distinguish genuinely representational states from non-representational states. The Organism-Receptor account does not entail that representations exist relative to construals, only that they are *ascribed* relative to construals. My account is consistent with the existence of a first-order account of the metaphysics of representation that justifies this practice (or doesn't). After all, the duck-rabbit can be construed as a duck even if it is not a duck, and nothing about that fact entails that ducks (or unambiguous images of ducks) are not real. The Organism-Receptor account describes a norm that plausibly guides human scientists with imperfect capacities for knowledge. But while my solution to the metadiscursive job description challenge is not inconsistent with Ramsey's solution to

the first-order job description challenge, it is inconsistent with Ramsey's characterization of scientific norms for ascribing natural representations.

7. Conclusion. I began by observing the common worry that scientists ascribe representations more liberally than many philosophers are comfortable with, and in particular that scientists rely on an unsatisfactory "receptor" criterion. I sketched an account on which scientists ascribe natural representations only to components of mechanisms of systems construed as "organisms." Since in practice cognitive scientists attend almost exclusively to systems that are easily so construed, their behavior may appear to be guided by the crude receptor criterion whereas in fact it is guided by the Organism-Receptor criterion. However, while the Organism-Receptor account is still relatively liberal, a crucial difference between the two accounts is that the crude criterion has absurd consequences, whereas such consequences are eliminated or marginalized on the Organism-Receptor criterion. Since scientists do not in fact endorse these absurd consequences, I argue that the augmented criterion is a better hypothesis regarding norms for representation-ascription in cognitive science.

This is proposal is not a comprehensive, new theory of representation, but it accomplishes two things. First, it provides argumentative resources for resisting the common worry that cognitive scientists use hopelessly liberal criteria for ascribing representations. Second, it offers a novel picture of practices for representation-ascription in the biological and behavioral sciences, one that is less pessimistic picture than Ramsey regarding conceptual rigor in cognitive science. The picture is not beyond criticism—in particular, it wants for a more detailed account of the grounds that warrant attributing behaviors and goals to systems. But since it is more faithful to our practice than Ramsey's it is likely to yield more productive suggestions for how to guide that practice into the future. I suggest that we safeguard conceptual rigor in cognitive science not by cleaving more faithfully to the representationalism of the

cognitive revolution, but by embracing role of construal in scientific inquiry, making it explicit, and subjecting it to reasoned criticism.

REFERENCES

Adams, Fred, and Ken Aizawa. 2001. "The Bounds of Cognition." *Philosophical Psychology* 14:43–64.

Adams, Fred, and Rebecca Garrison. 2013. "The Mark of the Cognitive." *Minds and Machines* 23:339–52.

Akagi, Mikio. 2017. "Rethinking the Problem of Cognition." *Synthese*. doi: 10.1007/s11229-017-1383-2.

Bechtel, William, and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421–41.

Brooks, Rodney. 1991. "Intelligence without Representation." *Artificial Intelligence* 47:139–59.

Brooks, Rodney, Jonathan Connell, and Peter Ning. 1988. "Herbert: A Second Generation Mobile Robot." *A.I. Memos* 1016:0–10.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58:7–19.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Cummins, Robert, and Pierre Poirier. 2004. "Representation and Indication." In *Representation in Mind: New Approaches to Mental Representation*, E sds. Hugh Clapin, Phillip Staines and Peter Slezak, 21–40. Amsterdam: Elsevier.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* 60:685–700.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

———. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT/Bradford.

Giere, Ronald N. 2006. *Scientific Perspectivism*. Chicago: University of Chicago Press.

Hubel, David H., and Torsten N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160:106–54.

Jastrow, Joseph. 1899. "The Mind's Eye." *Popular Science Monthly* 54:299–312.

MacFarlane, John. 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Clarendon.

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67:1–25.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

———. 1995. "Pushmi-Pullyu Representations." *Philosophical Perspectives* 9:185–200.

Piccinini, Gualtiero. 2015. *Physical Computation: A Mechanist Account*. Oxford: Oxford University Press.

Ramsey, William M. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.

Roberts, Robert C. 1988. "What Emotion Is: A Sketch." *Philosophical Review* 97:183–209.

Rowlands, Mark. 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: MIT Press.

Rupert, Robert. 2009. *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.

Seyfarth, Robert M., Dorothy L. Cheney, and Peter Marler. 1980. "Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication." *Science* 210:801–3.

Shea, Nicholas. 2013. "Naturalising Representational Content." *Philosophy Compass* 8:496–509.

Stinson, Catherine. 2016. "Mechanisms in Psychology: Ripping Nature at Its Seams." *Synthese* 193:1585–614.

Weiskopf, Daniel A. 2011. "Models and Mechanisms in Psychological Explanation." *Synthese* 183:313–38.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. 3rd Ed. Trans. G.E.M. Anscombe. Eds. G.E.M. Anscombe and Rush Rhees. Oxford: Blackwell, 2001.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -122-

Comparing Systems Without Single Language Privileging

Max Bialek

mbialek@rutgers.edu

For the 2018 PSA Meeting.

Word count: 4753

Abstract

It is a standard feature of the BSA and its variants that systematizations of the world competing to be the best must be expressed in the same language. This paper argues that such single language privileging is problematic because (1) it enhances the objection that the BSA is insufficiently objective, and (2) it breaks the parallel between the BSA and scientific practice by not letting laws and basic kinds be identified/discovered together. A solution to these problems and the ones that prompt single language privileging is proposed in the form of privileging the best system competition(s).

1 Introduction

According to the Best Systems Analysis (BSA), the laws of nature are the theorems of the best systematization of the world—with 'best' standardly understood to mean the simplest and most informative (on balance). It is currently a standard feature of the BSA (since Lewis 1983) and its variants (Loewer 2007; Schrenk 2008; Cohen and Callender 2009) that a single language must be privileged as the language in which all systems competing to be the best will be expressed. Two problems have led these authors to adopt single language privileging: The first is the Trivial Systems Problem (TSP), according to which, in brief, allowing for suitably gerrymandered languages can guarantee that the "best" system will have axioms and theorems undeserving of the name "law" (see Lewis 1983 for its initial development). Language privileging provides a quick fix to the TSP as long as the privileged language is not among the suitably (and problematically) gerrymandered. The second is the Problem of Immanent Comparisons (PIC) suggested by Cohen and Callender (2009). The PIC takes it to be the case that there are only "immanent" measures for simplicity, strength, and their balance—that is, measures defined for only one language. With single language privileging, no two systems ever need to be compared when expressed in different languages, and so having to use only immanent measures is not an issue.

Though single language privileging solves these problems for the BSA and its variants, it creates new ones of its own. For one, the BSA is already often criticized for being insufficiently objective—because it is unclear that there is an objective answer to the question of what makes a system the best—and single language privileging has the potential to fuel those criticisms by requiring proponents of the BSA to say which

language gets privileged. Relativizing laws to languages (as in Schrenk 2008 and Cohen and Callender 2009) goes some way to resist such criticisms, but, as Bialek (2017) argues, relativity itself should be minimized (as much as scientific practice allows) when responding to those who employ the 'insufficiently objective' critique of the BSA. Another issue with language privileging—a version of which is suggested in a specific critique of Lewis (1983) by van Fraassen (1989), and is here newly generalized as an issue for *any* single language privinleging—is that it breaks the supposedly close connection in scientific practice between the discovery of the laws and the discovery of basic kinds.[1]

Both problems are, ultimately, overstated, and may be resolved not with single language privileging, but with the privileging of *classes* of languages. This addresses both of the issues just raised. For one, it restores the co-discovery of laws and basic kinds to the BSA by making the search for laws (via a best system competition conducted in the course of scientific practice) include a search through a class of languages for the one that yields the best system-language pair. It also helps to limit the degree to which laws may need to be relativized to language by reducing the problem of privileging a language (class) to the already present problem of choosing a measure of 'best'.

The outline of this paper is as follows. I begin, in Section 2, by laying out the PIC. In Section 3, I argue that the PIC ignores the existence of measures (illustrated by the

[1]Depending on the specific interests of the author, there has been talk of "basic kinds" (as in Cohen and Callender 2009), "fundamental kinds" (Loewer 2007), and "perfectly natural predicates" (Lewis 1983). These are progressively more restrictive ways of interpreting the predicates of a language that appear in the axioms of a best system expressed in that language. Throughout the paper I use the more general phrase "basic kinds", but nothing about that usage precludes a more restrictive reading.

Akaike Information Criterion) that, while not transcendent (since they cannot compare systems expressed in *any* two languages), are also not immanent (since they can compare systems expressed in *some* different languages). Being sensitive to the existence of such measures suggests a slightly different problem of *transcendent* measures, which may be resolved through privileging classes of languages. The problem for single language privileging of breaking the connection between the discovering laws and basic kinds is developed in Section 4, and its resolution via language-class privileging is demonstrated. In Section 5, I argue that the question of which language class to privilege is reducible to the question of which measure(s) of 'best' (simplicity, informativeness, etc.) should be used. Lastly, in Section 6, I note that the reducibility just introduced suggests a new solution to the TSP that is focused on choosing appropriate measures of 'best', with the conclusion being that none of the problems that have prompted language privileging actually require it for their resolution.

2 The Problem of Immanent Comparisons

The "Problem of Immanent Comparisons" (PIC) begins with an appeal in Cohen and Callender (2009) to a distinction in Quine between *immanent* and *transcendent* notions. Quine writes: "A notion is immanent when defined for a particular language; transcendent when directed to languages generally" (Quine 1970, p. 19). Measurements of simplicity, since they depend on the language in which a system is expressed, are taken by Cohen and Callender to be immanent in this Quinean sense. Strength, or informativeness, is similarly immanent, since it is assumed to depend on the expressive power of the language in which a system is expressed. And, to finish out the set, balance

is said to be immanent as well, since it will be a measure dependent on immanent measures of simplicity and strength. If two systems are competing to be the best and are expressed in different languages, then we would need transcendent measures of simplicity, strength, and balance, in order to implement the best system competition. But "there are too few (viz. no) transcendent measures" of simplicity, strength, and balance (Cohen and Callender 2009, p. 8). Cohen and Callender write that

> *Prima facie*, the realization that simplicity, strength, and balance are immanent rather than transcendent—what we'll call *the problem of immanent comparisons*—is a devastating blow to the [BSA and its variants]. For what counts as a law according to that view depends on what is a Best System; but the immanence of simplicity and strength undercut the possibility of intersystem comparisons, and therefore the very idea of something's being a Best System.
>
> (Cohen and Callender 2009, p. 6, emphasis in original)

The only solution to the PIC, since (supposedly) systems can only be compared when they are expressed in the same language, is to adopt single language privileging.

3 Neither Immanent nor Transcendent

The issue with the PIC is that it ignores the existence of a large middle ground of measures that are neither immanent nor transcendent. To start, let us examine the central claim of the PIC: that simplicity, strength, and balance must be immanent measures. In defense of the idea that simplicity is immanent, Cohen and Callender

(2009, p. 5) defer to Goodman (1954) by way of Loewer, who writes: "Simplicity, being partly syntactical, is sensitive to the language in which a theory is formulated" (Loewer 1996, p. 109). Loewer and Goodman are exactly right. Simplicity is language sensitive. For example, let us adopt a naive version of simplicity, $SimpC(-)$, that is measured by the number of characters it takes to express a sentence (including spaces and punctuation). Consider the following sentence.

This sentence is simple.

Its $SimpC$-simplicity is 24 characters. The same sentence in Dutch is

Deze zin is eenvoudig.

The sentence's $SimpC$-simplicity now is 22 characters. So the $SimpC$-simplicity of a sentence depends or is sensitive to the language in which the sentence is expressed. Does that language sensitivity mean that $SimpC$ is immanent? It depends on what is meant by being "defined for a particular language".

$SimpC$ is, in some sense, "defined for a particular language". Insofar as the measure gives conflicting results for a sentence expressed in different languages, it would be ill-defined if we took it to be directed at sentences irrespective of the language in which they are expressed. One way of dealing with this would be to think that we have a multitude of distinct simplicity measures: $SimpC_{\text{English}}(-)$, $SimpC_{\text{Dutch}}(-)$, and so on. But doing that disguises an important fact: each of these measures of simplicity is *the same measure*, just relativized to particular languages. Drawing our inspiration from the "package deal" of Loewer (2007)—in which the BSA holds its competition between system-language pairs (or packages)—we could just as easily deal with the language

sensitivity of *SimpC* by saying it is defined for sentence-language pairs. We don't need, then, different measures of simplicity. Just the one will do:

$$\begin{aligned} SimpC(\ulcorner\text{This sentence is simple.}\urcorner, \text{English}) &= 24 \text{ char.} \\ SimpC(\ulcorner\text{This sentence is simple.}\urcorner, \text{Dutch}) &= 22 \text{ char.} \end{aligned}$$

In this way, *SimpC* is better understood as transcendent, and not immanent, because it is, as Quine put it, "directed to languages generally".

Of course, *SimpC* can't be directed to *all* languages, since it will be undefined for any languages that don't have a written form with discrete characters. This suggest that there is an important middle ground between immanent and transcendent measures. When a measure falls in that middle, as *SimpC* seems to, I will say that it is a "moderate measure".

So which conception of *SimpC* is the right one? The "devastating blow" that immanence deals to the BSA and its variants is that it "undercut[s] the possibility of intersystem comparisons" (Cohen and Callender 2009, p. 6). In our naive example,

$$SimpC_{\text{English}}(\ulcorner\text{This sentence is simple.}\urcorner)$$

is—if *SimpC* is immanent—incomparable to

$$SimpC_{\text{Dutch}}(\ulcorner\text{This sentence is simple.}\urcorner).$$

But obviously it's not. ⌜This sentence is simple.⌝ is *SimpC*-simpler in Dutch than in English (when being *SimpC*-simpler means having a lower value of *SimpC*).

Nothing prevents a transcendent or moderate measure from taking a language as one of its arguments. Such a measure is transcendent (or moderate), but language sensitive, and, importantly, it allows for comparisons even when a variety of languages are involved. That being the case, the mere language sensitivity of simplicity, strength, and their balance is not enough to guarantee that they are immanent, nor is it enough to guarantee the incomparability of systems expressed in different languages.

In response to the existence of a measure like *SimpC*, it might be suggested that there may well be transcendent (or moderate) measures plausibly named "simplicity" (etc.), but these are not the ones relevant to the BSA; the measures that *do* appear in BSA will be immanent. It is absolutely right to question the plausibility of a measure as naive as *SimpC* having a role to play in the BSA. (I certainly do not intend to defend *SimpC* as the right measure of simplicity for the BSA.) But I do not think it is clear why we should assume that the right measures are immanent. Rather, I think that moderate measures are, if anything, the norm, and an example may be found in the selection of statistical models.

Following Forster and Sober (1994), statistical model selection has standardly been associated in philosophy with the Akaike Information Criterion (AIC):

$$AIC(M) \;=\; 2[\text{number of parameters of } M] - 2[\text{maximum log-likelihood of } M]$$

The full details of AIC are not terribly important for our purposes here; it is enough to point out that that first term is concerned with the *number of parameters* of the statistical model M. Forster and Sober note that the number of parameters "is not a merely linguistic feature" of models Forster and Sober (1994, p. 9, fn. 13). But the

number of parameters is *a* linguistic feature of a model. Since AIC can compare models with different numbers of parameters, it can—if we think of statistical models as the system-language pairs of the BSA, and AIC as central to the best system competition[2]—compare systems expressed in different languages. AIC is thus a moderate measure.

It is important to note, however, that AIC is also not a transcendent measure. Kieseppä (2001) offers a response to critics of AIC who are concerned that the measure is sensitive to changing the number of parameters of a model by changing the model's linguistic representation. The response turns on the justification of "Rule-AIC", which says to pick the model with the smallest value of AIC, on the grounds that the predictive accuracy of model M is approximately the expected value of the maximum log-likelihood of M minus the number of parameters of M. Crucially,

> the theoretical justification of using (Rule-AIC) is valid when the considered models are such that the approximation [just mentioned] is a good one.
>
> (Kieseppä 2001, p. 775)

Let M be parameterized to have either k or k' parameters. Then there are two claims that are relevant to the justification of Rule-AIC:

$$\text{predictive accuracy of } M \approx E[(\text{maximum log-likelihood of } M) - k]$$

$$\text{predictive accuracy of } M \approx E[(\text{maximum log-likelihood of } M) - k']$$

[2]To make the connection between AIC and the BSA even stronger, it it worth noting that Forster and Sober (1994) take the "number of parameters" term to be tracking the simplicity of a model.

The predictive accuracy of M is independent of the number of parameters used to express M.[3] But the right side of the approximation in each claim *does* depend on the number of parameters. In general, both of these claims will not be true. Since Rule-AIC is only justified by the truth of these approximations, it will only be applicable to whichever parameterization of M makes the approximation true. The only time when both claims are true, and thus when AIC is applicable to both parameterizations, is when the difference between $E[(\text{maximum log-likelihood of } M) - k]$ and $E[(\text{maximum log-likelihood of } M) - k']$ is negligible. Kieseppä concludes:

> This simple argument shows once and for all that the fact that the number of the parameters of a model can be changed with a reparameterisation does not in any interesting sense make the results yielded by (Rule-AIC) dependent on the linguistic representation of the considered models.
>
> (Kieseppä 2001, p. 776)

From the epistemic perspective that is Kieseppä's concern, I can find room to agree that there is no "interesting sense" in which Rule-AIC is language dependent. This is because, if we are looking to employ Rule-AIC in statistical model selection, what is available to us is a procedure to check if the given parameterization is one that can support the justification of Rule-AIC. If the justification will work, then Rule-AIC applies, and if not, not. Rule-AIC isn't language dependent "in any interesting sense" insofar as it simply doesn't apply to the problematic languages/parameterizations that undermine its justification.

[3]This is intuitively true. It is also true in the formal definition of predictive accuracy given in Kieseppä (1997) and used in this argument from Kieseppä (2001).

However, from the perspective of the BSA and the PIC, these failures of Rule-AIC *are* interesting. AIC (the measure) is not immanent, but it is also not transcendent; it is merely moderate. *Some* reparameterizations of considered models will lead to the inapplicability of Rule-AIC. If Rule-AIC was how we were deciding which system was best, the existence of these problematic reparamterizations would be, as Cohen and Callender put it, a *prima facie* devastating blow to the BSA.

Towards the end of their introducing the PIC, Cohen and Callender write that

> What is needed to solve the problem is a *transcendent* simplicity/strength/balance comparison of each axiomatization against others. The problem is not that there are too many immanent measures and nothing to choose between them, but that there are too few (viz., no) transcendent measures.
>
> (Cohen and Callender 2009, p. 8, emphasis in original)

Cohen and Callender are probably right that there are "too few (viz., no) transcendent measures". In response to this, PIC says that measuring the goodness of a system must be done with immanent measures, and so no systems expressed in different languages may be compared in the best system competition. But non-transcendence is not a guarantee of immanence. We might call the problem that remains the *problem of transcendent measures* (PTC). Measures like AIC are not immanent, but they also aren't transcendent. That non-transcendence gives rise to a degree of language sensitivity that will *sometimes* prevent us from comparing systems expressed in different languages.

In response to the PIC and the supposed immanence of measures appropriate for the BSA, Cohen and Callender (2009) proposed the Better Best Systems Analysis (BBSA),

which relativizes laws to single languages. According to the BBSA, a best system competition is run for every language L (with some restrictions on "every" that aren't especially important here) where all the competing systems are expressed in L and the theorems of the system that is the victor of the competition are the laws *relative to* L. But now it seems that we might have at our and the BSA's disposal moderate measures. In the face of the non-transcendence of these measures—that is, in the face of the PTC—the BBSA's strategy of language relativity is still a good one.[4] Our language relativity does not, however, have to involve privileging *single* languages. The alternative is to relativize to *classes* of languages constructed to ensure the applicability of the measures employed in our best system competition.

4 Discovering Laws and Kinds Together

Before saying more about what relativizing laws to classes of languages would be like in any detail, it is important to say something about why we should pursue language-class relativity over the single language relativity of the BBSA. So, why should we? The reason is that one of the great virtues of the BSA and its variants is their offering of a metaphysics for laws that parallels the search for laws that is to be found in scientific practice, and that parallel is broken by single language privileging. A feature of the

[4]Without going into excessive detail about benefits (and costs) of the BBSA's relativity strategy over competitors, I hope it is enough to note that relativizing the laws allows us to sidestep the question of which language should be privileged entirely, since, ultimately, all languages will get a turn at being privileged, and thus, effectively, none are privileged over all.

search for laws in scientific practice is that it happens in conjunction with a search for the basic kinds of the world. This feature encourages us to acknowledge the importance of language in the BSA, since the basic kinds of the world are, presumably, going to correspond with the basic kinds that appear in the language in which the laws are expressed. Thus, when Lewis first recognizes the language sensitivity of simplicity, he concludes on a celebratory note by saying that the variant of single language privileging he introduces has the virtue of "explaining" why "laws and natural properties get discovered together" (Lewis 1983, p. 368).

For Loewer's Package Deal Analysis, the idea that laws and kinds are discovered together is central to the view. Indeed, the phrase "package deal" has its roots in Lewis, who says just before the "discovered together" remark that "the scientific investigation of laws and of natural properties is a package deal" (Lewis 1983, p. 368). While Loewer ultimately endorses a version of single language privileging, it is accompanied with a rough account of how a "final theory"—i.e., a candidate system-language pair—is arrived at:

> a final theory is evaluated with respect to, among the other virtues, the extent to which it is informative and explanatory about truths of scientific interest as formulated in [the present language of science] SL or any language $SL+$ that may succeed SL in the rational development of the sciences. By 'rational development' I mean developments that are considered within the scientific community to increase the simplicity, coherence, informativeness, explanatoriness, and other scientific virtues of a theory.
>
> (Loewer 2007, p. 325)

If the practice of science parallels the Package Deal Analysis, then the processes of discovering the laws and basic kinds are one and the same.

And it seems Cohen and Callender are also on board with laws and kinds being discovered together when they offer this nice remark on the phenomenon:

> historical disputes between theorists favoring very different choices of kinds seem to us to be disputes between two different sets of laws [...] it has happened in the history of science that people have objected to particular carvings—most famously, consider the outrage inspired by Newton's category of gravity. But given the link between laws and kinds, this outrage is probably best seen as an expression of the view that another System is Best, one without the offending category. If that other system doesn't in fact fare so well in the best system competition—as in the case of the systems proposed by Newton's foes—then the predictive strength and explanatory power of a putative Best System typically will win people over to the categorization employed. While it's true that some choices of [kinds] may strike us as odd, no one would accuse science—the enterprise that gives us entropy, dark energy, and charm—as conforming to pre-theoretic intuitions about the natural kinds of the world. Yet these odd kinds are all embedded in systematizations that would produce what we would consider laws.
>
> (Cohen and Callender 2009, pp. 17–18)

With everyone in agreement, what is the problem? Language privileging, essentially, happens *before* the identification (in the BSA and its variants) or discovery (in scientific practice) of the laws. Though Cohen and Callender will not "accuse science" of

"conforming to pre-theoretic intuitions about the natural kinds of the world", that is exactly what the BBSA (and any other single language privileging variant of the BSA) does when it privileges sets of kinds prior to a best system competition. Furthermore, PIC makes it such that "the predictive strength and explanatory power of a putative Best System" cannot "win people over to the categorization employed" because comparing two putative Best Systems expressed in different languages (with different "categorizations") is supposed to be impossible.[5]

Relativizing to classes of languages solves this problem. Scientists are able to approach the discovery of laws and kinds with pre-theoretic intuitions about how to systematize the world, the language to use when doing that, and the best system competition. As we will see below, the intuitions regarding language and the best system competition will locate them in a particular language class. Scientists will move away from their intuitions about language (and systematizing) when, much as Loewer describes above, there are languages in the relevant language class that may be paired with systems to yield a system-language pair that is scored better by the best system competition than the pre-theoretic system-language pair.[6]

[5]At least, it is impossible according to PIC for the BSA and its variants. If it *is* possible for scientists, then it is wholly unclear why it would be impossible for the BSA.

[6]This movement is only metaphorical for the BSA, where all the possibilities are considered and judged simultaneously. It is helpful, though, to think in the more methodical terms—of considering particular transitions from one system-language pair to another, the benefits that they might bring, and then adopting them or not—because that is what will happen in actual scientific practice.

5 Limiting Language Relativity

Let us begin addressing how language-class relativity can work by looking in more detail at the single language relativity of the BBSA. In the BBSA, there are the fundamental kinds K_{fund}. The set of all kinds $\mathcal{K}$ is the set including K_{fund} closed with respect to supervenience relations—that is, $\mathcal{K}$ includes every kind that can be defined as supervening on the arrangement of the K_{fund} kinds in the actual world. A language L is determined by the set of kinds for which it has basic predicates, and there is a language L_i for every $K_i \subseteq \mathcal{K}$. For any two languages L and L', the supervenience relations between the kinds of the languages and K_{fund} can be thought of as schemes for *translation* between L and L'. The set of all languages $\mathcal{L}_{\text{all}}$ can be thought of as the set of languages that includes L_{fund} closed with respect to all translations. A class of languages $\mathcal{L}_i$ is a set of languages including L_{fund} closed with respect to some acceptable (all, in the case of $\mathcal{L}_{\text{all}}$) translations.

To illustrate, let us consider a 'coin flip' world. Such a world is a string of `H`s and `T`s, which we will assume are the only two fundamental kinds. Another set of kinds might be $K_{\text{ex}} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$, where the translation that gets us to the corresponding language L_{ex} from L_{fund} maps the pairs `HH`, `HT`, `TH`, and `TT`, to `a` through `d`, respectively. An example of a class of languages that includes L_{ex} could be $\mathcal{L}_{\text{n-tuple}}$: Let an acceptable translation for $\mathcal{L}_{n\text{-tuple}}$ be one that, for a given n takes the set of all n-tuples of `H` and `T`, and maps them to a set of kinds $K_n = \{k_{n,1}, k_{n,2}, ...k_{n,2^n}\}$. L_{fund}, then, is just L_1. When `a` through `d` are $k_{2,1}$ through $k_{2,4}$, our K_{ex} and L_{ex} are precisely K_2 and L_2. All, and only, the languages that may be formed through this procedure will be members of the class $\mathcal{L}_{\text{n-tuple}}$.

A language-class relative variant of the BSA will run a best system competition for

every class of languages $\mathcal{L}_i$. Then $\mathcal{S}$ is the set of all systematizations of the world, the set of all competing system-language pairs for the $\mathcal{L}_i$-relative best system competition is given by $\mathcal{S} \times \mathcal{L}_i$.

We can apply this conception of language-class relativity to our other running example of statistical model selection with AIC. Recall that *some* reparameterizations of statistical models would prove problematic for the use of AIC. To reparameterize a model is akin to translating it from one language to another. We can understand, then, the problem of language sensitivity for AIC as being related to some set of problematic translations. If we subtract these problematic translations from the set of all translations, then we have a set of acceptable translations which defines a class of languages that we can call $\mathcal{L}_{AIC}$. $\mathcal{L}_{AIC}$ is precisely the set of all languages such that a system expressed in any one of them will be comparable to a system expressed in any other using AIC. As long as the moderate measures used in the best system competition have clearly problematic and/or acceptable translations associated with them, then the class of languages that may be used to express competing systems will be determined by the measures used in the best system competition.

This will have one of two effects on the extent to which the BSA must be relativized to classes of languages, but before going into those details it will be helpful to characterize "competition relativity". Competition relativity should be understood in much the same way that language relativity is understood. The competition of the BSA is the thing that takes system-language pairs as its inputs, and outputs a best pair from which we can read off the laws. The competition decides what system-language pair is best by considering how well they measure up with respect to some collection of theoretical virtues (like simplicity and informativeness) and the actual world. Much as

we might worry about what language to privilege, and side-step that problem by relativizing laws to languages so that every language takes a turn as the privileged one, we might also worry about which competition, or which set of theoretical virtues, to privilege. Competition relativity sidesteps the problem of which collection of theoretical virtues to use (and weighting between them, and means of measuring them, etc.) by relativizing laws to every way of formulating a best system competition.[7]

So, either the BSA will be committed to competition relativity or not. Suppose that it is not. For convenience, suppose further that Rule-AIC is all that there is to the best system competition. In that case, the BSA will always be run using the $\mathcal{L}_{AIC}$ class of languages. Language-class relativity is not required since there is only one language class that will ever be relevant to the BSA—namely $\mathcal{L}_{AIC}$, as determined by the best system competition. Now suppose that there is competition relativity. A different best system competition must be run for every competition function C_i in the set of all possible competition functions $\mathcal{C}$. In principle we will need to run best systems competitions for every pair in $\mathcal{C} \times \mathbb{L}$, where $\mathbb{L}$ is the set of all language classes. Let $\mathcal{L}_j$ be the class of languages constructed according to the translations that are acceptable for the measures that comprise C_i when $i = j$. In practice, however, it will only make sense to run a competition once for each $C_i \in \mathcal{C}$, since the pairs $C_i, \mathcal{L}_j$ will be unproblematic only when $i = j$. Language-class relativity in this situation will be redundant with competition relativity. We also have it that, in either case (of needing competition relativity or not), single language relativity remains unnecessary for all the same reasons that recommended language-class relativity.

[7]See Bialek (2017) for an extended discussion of competition relativity and the possibility of its inclusion in the BSA.

6 The Trivial Systems Problem

The redundancy of any sort of language privileging relativity with competition relativity offers an interesting solution to the Trivial Systems Problem (TSP) that initiated the trend of single language privileging.

Recall that the TSP is concerned with the possibility of suitably gerrymandered languages that can guarantee that the "best" system will have axioms and theorems undeserving of the name "law". In the introduction fo the problem, Lewis imagines a system S and predicate F "that applies to all and only things at worlds where S holds" (Lewis 1983, p. 367). The system S, then, maybe be expressed by the single axiom $\forall xFx$, simultaneously achieving incredible informativeness—because of the specific applicability of F—and incredible simplicity—because, Lewis assumes, '$\forall xFx$' is about as simple as a system could be. So S will be the best system despite a variety of reasons why it shouldn't be, the foremost of which are that: (1) $\forall xFx$ will be a law unlike any we would expect to find, (2) F would be a basic kind unlike any we would expect to find, and (3) every regularity of the world is a theorem of $\forall xFx$, so there would be no distinction between accidental and lawful regularities.

The problem is solved as long as we can avoid languages that include problematic predicates like F. Single language privileging solves this problem as long as the privileged language does not include the (or any) problematic predicate(s). Language-class privileging likewise solves the problem as long as no language in the class includes the (or any) problematic predicate(s). That alone might be enough said, but the redundancy of language-class choice on competition choice offers a more nuanced solution: The best system competition could be chosen such that the corresponding class

of languages does not include F or any similarly problematic predicates. But it could also be chosen such that F and its ilk are certain to not be the best. Lewis assumes with no discussion that $\forall xFx$ is an incredibly informative and simple system, but, even if that is true for the measures/competition, it need not be true for every competition. If there is competition relativity, then there may be competitions for which a trivial system like $\forall xFx$ is the victor, but for the same reasons that such a system is problematic, scientists will simply be uninterested in the laws relative to those competitions.[8] If there isn't competition relativity, it seems unlikely that science would unequivocally endorse a competition that yields a trivial system (or, if it does, then we would need to take a step back and seriously reconsider our aversion to such a system).

In the end, there is no apparent need for any language privileging or relativity in the BSA.[9] Its role in solving the problems of immanent (or transcendent) comparisons and trivial systems will be unnecessary (if a single moderate best system competition can be identified) or redundant with competition relativity.

[8]In much the same way that Cohen and Callender (2009) allow for there to be uninteresting sets of laws determined relative to languages that include F-like predicates.

[9]The problems discussed is not the only reason one might want to adopt language relativity in the BSA. It should also be noted that one of the virtues of the BBSA's single language relativity is that it allows the view to accommodate an egalitarian conception of special science laws. Language relativity, however, is not the only way of getting special science laws out of the BSA. This is an important issue to which the discussion in this paper is relevant, but a proper exploration of it warrants a more focused and extended treatment.

References

Bialek, M. (2017). Interest relativism in the best system analysis of laws. *Synthese 194*(12), 4643–4655.

Cohen, J. and C. Callender (2009). A better best system account of lawhood. *Philosophical Studies 145*(1), 1–34.

Forster, M. and E. Sober (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science 45*(1), 1–35.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Kieseppä, I. (1997). Akaike information criterion, curve-fitting, and the philosophical problem of simplicity. *The British journal for the philosophy of science 48*(1), 21–48.

Kieseppä, I. (2001). Statistical model selection criteria and the philosophical problem of underdetermination. *The British journal for the philosophy of science 52*(4), 761–794.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343–377.

Loewer, B. (1996). Humean supervenience. *Philosophical Topics 24*(1), 101–127.

Loewer, B. (2007). Laws and natural properties. *Philosophical Topics 35*(1/2), 313–328.

Quine, W. V. O. (1970). *Philosophy of logic*. Harvard University Press.

Schrenk, M. (2008). A theory for special science laws. In S. W. H. Bohse, K. Dreimann (Ed.), *Selected Papers Contributed to the Sections of GAP.6*, pp. 121–131. Paderborn: Mentis.

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.

Explaining Scientific Collaboration: a General Functional Account

Thomas Boyer-Kassem* and Cyrille Imbert†

October, 2018

Abstract

For two centuries, collaborative research has become increasingly widespread. Various explanations of this trend have been proposed. Here, we offer a novel functional explanation of it. It differs from accounts like that of Wray (2002) by the precise socio-epistemic mechanism that grounds the beneficialness of collaboration. Boyer-Kassem and Imbert (2015) show how minor differences in the step-efficiency of collaborative groups can make them much more successful in particular configurations. We investigate this model further, derive robust social patterns concerning the general successfulness of collaborative groups, and argue that these patterns can be used to defend a general functional account.

*MAPP (EA 2626), Univ. Poitiers, France. thomas.boyer.kassem@univ-poitiers.fr
†CNRS, Archives Poincaré, France. cyrille.imbert@univ-lorraine.fr

1 Introduction

For two centuries, co-authoring papers has become increasingly widespread in academia (Price, 1963, Beaver and Rosen, 1979), especially in the last few decades. Since the 1950s, the percentage of co-authored papers has grown at a common rhythm for science and engineering, social sciences, and patents; the mean size of collaborative teams has also increased, and even more so in science and engineering. No such increase is visible for the art and humanities (Wuchty et alii, 2007).

Various explanations of this collaborative trend have been proposed: for example, it may be caused by scientific specialization, it may increase the productivity or reliability of researchers, or be promoted by the rules of credit attribution. Here, we aim at offering a new functional explanation of this trend by showing that collaboration exists because it increases the successfulness of scientists. The present explanation differs from accounts like that of Wray (2002) by the social and epistemic mechanism that grounds the beneficialness of collaboration. We analyze further an existing model that shows how minor differences in the step-efficiency of collaborative groups at passing the steps of a project can make them much more successful in particular configurations (Boyer-Kassem and Imbert, 2015) and show how it can be used to build a general and robust functional explanation of collaboration.

We introduce the model in section 2. After presenting functional explanations (section 3), we show how the model can be used to derive robust social patterns of the successfulness of collaborative groups (section 4), and argue that these patterns can refine and strengthen functional explanations of collaboration like the one defended by Wray (sections 5 and 6).

2 Boyer-Kassem and Imbert's Model: Main Results and Explanatory Lacunas

Boyer-Kassem and Imbert (2015) investigate a model in which n agents struggle over the completion of a research project composed of l sequential steps. At each time interval, agents have independent probabilities p of passing a step. When an agent reaches the end of the project, she wins all the scientific credit and the race stops (this is the priority rule). Agents can organize themselves into collaborative groups for the whole project, meaning that they only share information, i.e. step discoveries — clearly, there are more favorable hypotheses associated with collaborating, like having new ideas or double-checking (see below). Within a group, agents make progress together, and equally share final rewards. Thus, a group of k agents (hereafter k-group) passes a step with probability $p_g(k,p) = 1-(1-p)^k$. In forthcoming illustra-

tions, the value of l is set to 10 and that of p to 0.5, which is not particularly favorable for groups (ibidem, 674). If collaboration is beneficial with these hypotheses, it will be even more so with more favorable or realistic ones. A community of n agents (hereafter, n-community) can be organized in various k-groups. For example, a 3-community can correspond to configurations (1-1-1), (2-1) or (3). The individual successfulness of an agent in a k-group in a particular configuration is defined as the average individual reward divided by time. It has been obtained for all configurations up to $n = 10$, on millions of runs.

Note that this model is not aimed at quantifying the actual successfulness of collaborative agents, but at analyzing the differential successfulness of agents depending on their collaborative behavior. The main finding is that minor differences in the efficiency at passing steps can be much amplified and that, even with not-so-favorable hypotheses, collaboration can be extremely beneficial for scientists. For example, in a (5-4) (resp. (2-1)) configuration, whereas the difference in step efficiency between the 5 (resp. 2) and the 4-group (resp. 1-group) is 3% (resp. 50%), the difference in individual successfulness is 25% (resp. 700%). The scope of these results actually goes beyond the initial hypotheses in terms of information sharing. Formally speaking, the model is a race between (collective) agents i with probabilities p_i of passing steps. *Whatever the origin* of the differences in p_i, they are greatly amplified by the sequential race. In other words, any factor, whether epistemic or not, that implies an increase in p_i of a k-group (e.g. if a collaborator is an expert concerning specific steps, if increased resources improve step-efficiency, etc.) makes this group as successful as a larger group — hence the generality of this mechanism.

Still, these results do not explain scientific collaboration by themselves. First, collaboration is beneficial for particular k-groups in particular configurations only: a 2-group is very successful in configuration (2-1-1-1-1) but not in (7-2). Thus, the model mostly provides possibility results about what can be the case in certain configurations. Second, the explanandum is a general social feature of modern science, not some collaborative behavior in some particular case, so the explanans must also involve general statements about the link between collaboration and beneficialness. Then, if the model presents generic social mechanisms with explanatory import, one needs to describe at a general level the effects of these mechanisms and provide some general, invariant pattern between collaboration and beneficialness. This is what we do in section 4. A final serious worry is that the beneficialness of a state by no means explains why it exists, nor perseveres in being. A link needs to be made between the beneficialness of collaboration and its existence over time. We suggest that this connection can be accounted for functionally.

3 Functional Explanations and Collaboration

We review in this section how functional explanations work and how they can be used in the present case. We follow Wray's choice to use Kincaid's account because it is simple, widely accepted, and that nothing substantial hinges on this choice. Functional explanations explain the existence of a feature by one of its effects, usually its usefulness or beneficialness. As such, they can be sloppy and badly flawed. The usefulness of the nose to carry glasses does not explain that humans have one. Nevertheless, if stringent conditions are met, it is usually considered that functional explanations can be satisfactory, typically within biology. Even Elster, who otherwise favors methodological individualism, agrees that functional explanations can be acceptable in the social science (Elster, 1983). According to Kincaid (1996, 105-114), P is functionally explained by E, i.e. P exists "in order to promote <effect E>" if:

(1) P causes E,

(2) P persists because it causes E,

(3) P is causally prior to E.

Then, a functional explanation of collaboration should have the following form:

(1c) Scientists' collaborative behavior causes the increase of their individual successfulness.

(2c) Scientists' collaborative behavior persists (or develops) because it causes a higher individual successfulness.

(3c) Collaborative behavior is causally prior to this increased individual successfulness that is rooted in collaborative behavior.

We agree with Wray (2002, 161) that it is implausible to consider that the high successfulness of scientists is the initial cause of collaboration since many scientists have been successful (and continue to be in some fields) without collaborating. In the same time, there can be various contingent reasons why some researchers have decided to engage in some collaboration. So, what calls for an explanation is the fact that collaboration is widespread and persistent, not its occasional existence.

4 Collaboration Causes Successfulness

We now argue that the above model provides strong evidence in favor of (1c). To explain the general collaborative patterns described above, the causal

relation between collaboration and successfulness needs to be general and robust. Hence, one needs to go beyond the description of the beneficialness of collaboration in particular situations. A first route is to find general results about when it is beneficial for individuals to collaborate, such as the following theorem (see the appendix for the proof).

Theorem. When m groups of equal size k merge, the individual successfulness of agents increases.

In other words, as soon as several k-groups of the same size exist, they would improve the individual successfulness of their members by merging. A corollary is that single individuals always have interest in collaborating. However, this theorem only covers a small subset of possible configurations, and cannot provide a general vindication for the causality claim (1c). Further, agents might only use it if they are aware of it and are in a position to identify groups of equal-size competitors, which cannot be assumed in general.

To overcome these difficulties, we now assess agents' successfulness irrespective of what they know about other competitors: we consider the average successfulness of k-groups over all possible configurations for each community size. For example, we average the individual successfulness of 4-groups in configurations (4-1-1-1); (4-2-1) and (4-3)[1]. In order to study the robustness of the causal relation between collaboration and successfulness, we investigate in the next paragraphs how much collaborating remains beneficial under variations of key parameters of the competition context.

Successfulness and community size. Figure 1 shows the average successfulness within k-groups for communities of various sizes. First, the successfulness of loners brutally collapses and is much lower than that of other k-groups as soon as $n > 2$. This confirms that except when nobody collaborates, or in very small communities, loners are outraced. Second, for all group sizes, individual successfulness decreases for larger communities, as can be expected when the number of competing groups and their size increases. Nevertheless, the successfulness of k-groups remains high and stable up to some community size s larger than k till they are eventually outperformed by larger groups or till growing bigger would mean over-collaborating (see (Boyer-Kassem and Imbert, 2015, 679-80) for an analysis of over-collaboration in large groups). Third, the larger the groups are, the longer and flatter this initial plate of successfulness is and the less steep the decrease in successfulness is. Fourth,

[1]There is no clear rationale about how to weigh configurations. From a combinatorial viewpoint, configuration (1,1,1,1,1,1,1) has one realization and (3,2,1,1) several ones. But from an empirical viewpoint, when scientists hardly collaborate, configuration (1,1,1,1,1,1,1) is usual and (3,2,1,1) extremely rare. We have privileged simplicity and chosen to give equal weight to all configurations.

when n is much larger than k, the successfulness of k-groups increases with k. However, this increase is a moderate one and small groups still do reasonably well, which is somewhat unexpected, given the general amplification effect — but see the analysis of figure 3 below for more refined analyses. Typically, in 10-communities, 2-groups do badly but remain somewhat viable since their average successfulness remains between 1/3 to 1/2 of that of 3 or 4-groups. Overall, not collaborating is in general not a viable strategy. Collaborating moderately ($k = 2$ or 3) can be very rewarding when there are few competitors (e.g. in small research communities, or on ground-breaking questions that are only known to a handful of scientits). Smalls groups remain viable but tend to be outraced when communities become significantly larger (typically, concerning questions belonging to normal science that many researchers are likely to tackle). Thus, moderately collaborating is a viable but more risky strategy when uncertainty prevails about the number and size of competing groups. Finally, while large collaborative groups rarely get exceptionally high gains, they are extremely safe, with moderate differences in successfulness between them or when the community size increases.

Successfulness and group size. Figure 2 shows the variation of individual successfulness with group size for various community sizes. First, for $n > 2$, the successfulness curve has a one-peaked (discrete) form, the maximum of which grows with the community size. Second, these one-peaked curves are not symmetric: the increase in successfulness is steep (but less so for larger groups), the decrease is gradual (idem). Large groups predate resources so groups need to grow big quickly to get some share and because returns can be increasing (Boyer-Kassem and Imbert 2015, 678), the increase in successfulness is steep. The decrease after the peak is slow because large group are hard to predate but over-collaborating can become suboptimal when the increase in gain by predation no longer makes up for the need to share between more people). These results are not trivial because at the configuration level, the successfulness of groups is contextual. They are important, too. A one-peaked profile is usually *assumed* in the literature about coalitions. Here, it emerges from a micro-model, and gets its justification from it. Overall, these patterns show again that agents have a large incentive to collaborate substantially, whatever the competing environment.

Successfulness in more or less collaborative communities. Figure 3 finally shows how the successfulness of k-group members varies with the degree of collaboration in their competition environment.[2] Here again, what matters

[2]Here, the degree of collaboration in each configuration is assessed by computing the average size of k-groups. For each k, we then compute the average successfulness of a member of a k-group over configurations having a degree of collaboration within intervals $[1, 1.5]$ (represented at coordinate "1.25" on the x-axis), $[1.25, 1.75]$, $[1.5, 2]$ ... $[3.5, 4]$. We

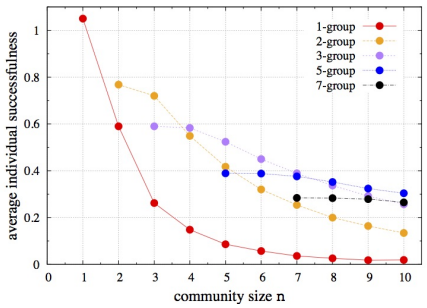


Figure 1: Variation of individual successfulness with community size.

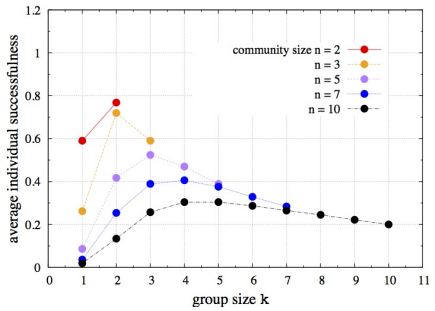


Figure 2: Variation of individual successfulness with the size of groups.

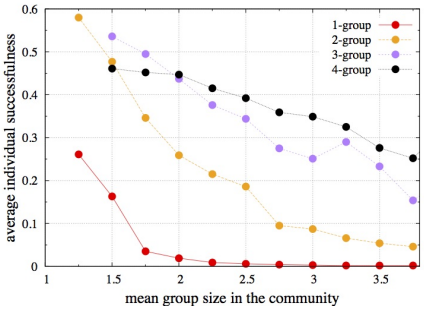


Figure 3: Variation of successfulness with the degree of collaboration in communities.

is less the exact value of the successfulness than the differential successfulness between more or less collaborating individuals. The graph confirms that successfulness depends less on the absolute size of groups than on how much they collaborate in comparison with their competitors. Scientists who collaborate more than average are very successful; those who collaborate as their peers do reasonably well; those that collaborate less than average are outraced by a large margin. This general result is not unexpected given all the above results, but the graph highlights that success for intensively collaborating scientists, and underachievement for under-collaborators can be very large. This is an important finding because if, as we shall see, successful scientists pass over their collaborative habits more than their peers, then the feedback loop provides a mechanism that favors the *increase* of the degree of collaboration by promoting those that collaborate more than others.

have chosen overlapping intervals to smoothen results. The average is computed up to communities of size 10.

Partial conclusion. Overall, the results show that — everything else being equal — collaborating a lot entails successfulness. This relation is robust under changes in the size of communities or in the exact size of groups. Further, those who collaborate more than average are much more successful. Collaborating too much is not a significant problem, under-collaborating is. So, collaborating a lot is a safe working habit, especially in the absence of information about the size and structure of the competing community. In light of this evidence, (1c) seems adequately supported.

5 Collaborative Practices Develop Because of the Success of Collaborative Scientists

We have so far argued that collaborative scientists, especially when they collaborate more than others, are more successful. We now need to argue that, because of this differential successfulness, collaborative habits persist and possibly develop in scientific communities (2c). A wide variety of social mechanisms across scientific contexts can contribute to this feedback loop. Accordingly, we shall be content with giving various evidence that strongly suggests that this link is a likely one.

Transmission. Knowing how and when to collaborate is not straightforward. Like other know-how skills, it can be developed by exercising it with people who already possess the relevant procedural knowledge. In this case, people who already collaborate can endorse this role of cultural transmission for colleagues and above all students (Thagard, 2006). Working with students is an efficient way to train them as scientists (Thagard, 1997, 248—50), so scientists have incentives to enroll students in their collaborative groups. Then, the cultural transmission of collaborative practice does not require any particular effort on top of that. The very circumstances that make collaboration possible and beneficial also make its transmission easier: when a research project can be divided into well-defined tasks, the solutions of which can be publicly assessed and shared, it is easier to enroll other people and thereby transmit collaborative skills to them (ibidem). Thus, collaborative habits can be passed over and need not be reinvented by newcomers.

Transmission opportunities. We now argue that collaborative scientists, because they are more successful, will more often be in a position to transmit their collaborative habits and that the collaboration rate will therefore increase. Within applied science, in which collaboration is also widespread (Wuchty, 2007), research projects are usually directed at finding profitable applications, which can be patented. Thus, fund providers are directly and strongly interested in hiring and providing resource to successful scientists,

who develop such applications. Within pure science, the connection is less straightforward. But because scientific success is the official goal of science, successful scientists can be expected to stand better chances to get good positions and grants, develop research programs, and pass over their collaborative habits.

Note that it is merely needed that the function between the pragmatic rewards of scientists and their success is on average increasing. This remains compatible with the fact that *some* epistemically successful scientists get little resource and *some* unsuccessful scientists get a lot — which seems to be the case. Actually, non-epistemic factors may even tend to over-credit successful scientists, and in particular collaborative ones. First, individual successfulness has been assessed in the model with a conservative estimate. It seems that an agent's publication within a k-group is actually more appreciated than just $1/k$ of a single-authored publication. For instance, a large French research institution in medicine officially weighs the citations of a paper with "a factor 1 for first or last author, 0.5 for second or next to last, and 0.25 for all others" (Inserm 2005). Also, a publication within a 10-group will generally be more visible than one single-authored publication, since more people can promote or publicize collective publications and research topics. Second, sociology of science seems to indicate that scientific credit tends to accrue to a subset of scientists who are perceived as extremely successful — this is the Matthew effect (Merton, 1968). Then, to the extent that access to resources increases with scientific credit, successful collaborative scientists can be expected to benefit from this effect and transmit more their working habits. The concentration of credit and resource may further stimulate collaborative behavior with these fortunate scientists.

Other types of mechanisms may contribute to this process, like conscious ones. So far, agents have only been supposed to follow their working habits and sometimes transmit them. But supplementary intentional or imitative processes may also feed this dynamics [3]. Once winners of the scientific race publish co-authored articles, it becomes easy for others to see that successful scientists are highly collaborative ones. (For instance, if agents of a 3-group are 4 times more successful than a single agent, this means that their groups publishes 12 more articles than this agent). Accordingly, the belief that collaborating is beneficial can be acquired as collaborating becomes usual. Furthermore, resources may accrue to scientific institutions that host individually successful scientists, and indirectly to these scientists. Agents in the model can be reinterpreted as teams or collective entities which decide to share results or to combine their expertise to produce collective articles. Then, these institutions

[3]Kincaid mentions that "complex combinations of intentional action, unintended consequences of intentional action, and differential survival of social practices might likewise make these conditions [(1)–(3) in our Section 3] true" (Kincaid 1996, 112).

and their members will be more successful, may attract resource, and will keep developing and transmitting their working habits.

In light of the above discussion, we believe that the causal connection between the success of collaborative scientists and the persistence and development of collaborative practices is highly plausible.

6 Discussion

Good functional explanations should be unambiguous about when the causal mechanisms that they rely on are efficient. In the present case, the following conditions can be emphasized.

First, conditions for the application of the priority rule should be met. In particular, (i) it should be possible to single out problems and to state uncontroversially when they are solved. Second, for the model to apply, (ii) scientific problems should be dividable into subtasks, and (iii) the solutions of these subtasks should be communicable. Finally, the model assumes that (iv) the completion of these subtasks should be sequential, but our conclusions still hold if this condition is relaxed. Indeed, if some subtasks can be tackled in parallel then the project can be completed even more quickly by different agents of a group, and collaboration is even more successful. Conditions (i)-(iii) are somewhat met in the formal and empirical sciences, less so in the social science, and almost not in the humanities. For example, as noted by Thagard (1997, 249), the humanities do not obviously lend themselves to the division of labor and to teacher/apprentice collaborations. Similarly, the importance of interpretative methods and the coexistence of incompatible traditions may prevent consensus on the nature of significant problems and what counts as a solution. This may account for the differences concerning collaborative patterns in these fields.

As mentioned above, different causal pathways may connect the successfulness of collaborative scientists to the persistence and development of collaborative practices. Thus, conditions for the fulfillment of claim (2c) cannot be uniquely specified. But several points are worth mentioning. First, the activity of epistemically successful scientists should be favored by scientific institutions. This can be the case if it is agreed that scientific success, in the form of publications or patents, is valued and promoted. Concerning scientific results that lead to patents, applications and financial gains, this condition is met when public or private funders value such outputs. Concerning pure scientific results, this means that there should be a wide agreement about which results are scientifically good and significant, and there should exist common and accessible publication venues, the value of which is consensual. Again, these conditions are approximately met in the formal and empirical sciences, less so in the social science and, almost not in the humanities in which scholars do not share paradigms, methods or norms about what is scientifically sound

and significant, and cultural and linguistic barriers can restrain the existence of unified communities and common publication venues. Second, in contexts in which researchers and projects are regularly evaluated, especially by agents or institutions who are not in a position to asses the scientific value of their work, the existence of a common standard of success in terms of publications (through simple and calibrated publication indicators) may even more favor researchers who are successful, and therefore the development of collaboration. Finally, when resources are crucial to carry out or facilitate research, snowball effects can favor even more successful scientists, and in particular collaborative ones. This resource accessibility condition, which is central in Wray's explanation, is not in ours. But we agree that in such cases, the functional mechanisms that we describe will be even stronger. In this sense, our account encompasses Wray's. This condition about resources may be another reason for the difference in collaborative behavior between the formal or empirical sciences, the social sciences and the humanities.

7 Conclusion

We have argued that collaborating a lot is overall a safe and success-conducing practice. This conclusion is robust for various sizes of groups, communities and degrees of collaboration; everything being equal, those who collaborate more than average do better. Then, to the extent that the successfulness of researchers gives them more opportunities to transmit their research habits, the development of collaborative practices in communities can be functionally explained. We have further emphasized that the conditions for this functional pattern to work are specifically met in the scientific fields in which collaboration is well-developed. Accordingly, it seems reasonable to consider that this functional mechanism is an important element of the explanation of the development of collaboration in modern science.

The explanation of collaboration is probably a multi-factorial issue. Nevertheless, an asset of our general functional explanation is that it highlights the unexpected force of beneficial aspects of collaborative activities and suggests important roles for contextual factors that are associated with the rise of collaboration. As such, it is general and unifying. For instance, the competition model shows how the division of scientific labor, the use of specialized experts (Muldoon 2017), or the increased reliability of collaborative teams (Fallis 2006, 200) can increase the probability that groups pass research steps and have amplified effects in terms of successfulness. Similarly, factors like the need to access resources to carry out or facilitate research can create a snowball effect that favors epistemically successful (collaborative) researchers (Wray 2002). And factors like the globalization of research or professionalization (Beaver, 1979) can be seen as conditions favoring the application of the priority rule

and scientific competition.

Finally, while nothing in the model provides an internal limit to the growth of collaboration, one can note that there is a wealth of reasons why collaborating groups cannot develop forever. For example, communities are limited in size, spatially distributed, and collaboration is all the more costly as groups are large. The model could be easily modified to integrate factors that limit the success and development of collaboration.

8 Appendix: Proof of the Theorem

Consider first the simple case where the m k-groups don't have other competitors. By symmetry, all groups have the same probability $1/m$ to win the race and get the reward — call this reward r. So, the individual expected reward is $r/(km)$. Suppose now the groups merge and all km agents collaborate. Each of them will receive the same reward, so their expected individual rewards are $r/(km)$ too. However, what matters in the model is not the expected reward, but the successfulness, which is this quantity divided by time. Because within a collaboration agents share all the steps they pass, the larger km-group will be at least as quick, and sometimes more, than all k-groups — more precisely: for a given drawing of all random variables corresponding to attempts to pass the steps, for all agents and temporal intervals, the km-group will move at least as quickly as all k-groups. So the individual successfulness is at least as high when identical groups merge.

Consider now the case where there are other competitors than the m groups. For a given drawing of all random variables, either the winner is one of the m groups, or another competitor. In the former case, the above reasoning can be made again, and the same conclusion holds. In the latter case, there is nothing to lose, and because the km-group is sometimes quicker than the m k-groups, there can be additional cases where it outcompetes the other competitors; then, the individual successfulness increases with the merging. QED.

9 References

Beaver, Donald deB. and Rosen, Richard (1979) "Studies in Scientific Collaboration: Part III", *Scientometrics*, 1(3): 231-245.

Boyer-Kassem, Thomas, and Cyrille Imbert (2015), "Scientific Collaboration: Do Two Heads Need to Be More than Twice Better than One?" Philosophy of Science 82 (4): 667–88.

Elster, Jon (1983), *Explaining Technical Change: A Case Study in the Philosophy of Science*, Studies in Rationality and Social Change, New York: Cambridge University Press.

Fallis, Don (2006), "The Epistemic Costs and Benefits of Collaboration", *Southern Journal of Philosophy* 44 S: 197–208.

INSERM (2005), "Les indicateurs bibliométriques à l'INSERM", `https://www.eva2.inserm.fr/EVA/jsp/Bibliometrie/Doc/Indicateurs/Indicateurs_bibliometriqu Inserm.pdf`

Kincaid, Harold (1996), *Philosophical Foundations of the Social Sciences*, Cambridge University Press.

Merton, Robert K. (1968), "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered", *Science*, 159 (3810): 56–63.

Muldoon, Ryan (2017), "Diversity, Rationality, and the Division of Cognitive Labor", in Boyer-Kassem, T., Mayo-Wilson, C. and Weisberg, M. (eds.), *Scientific Collaboration and Collective Knowledge*, New York: Oxford University Press.

Price, Derek John de Solla (1963), *Little Science, Big Science*, New York, Columbia University Press.

Thagard, Paul (1997), "Collaborative Knowledge", *Nous* 31(2): 242—261.

— (2006), "How to Collaborate: Procedural Knowledge in the Cooperative Development of Science", *The Southern Journal of Philosophy*, XLIV: 177—196.

Wray, K. Brad (2002), "The Epistemic Significance of Collaborative Research", *Philosophy of Science* 69 (1): 150-168.

Wuchty, Stefan, Jones, Benjamin F. and Uzzi, Brian (2007), "The Increasing Dominance of Teams in Production of Knowledge", *Science* 316(5827): 1036-1039.

Individuating Genes as Types or Individuals: Philosophical Implications on Individuality, Kinds, and Gene Concepts

Ruey-Lin Chen
Department of Philosophy
National Chung Cheng University
This paper will be presented at PSA 2018 meeting at Seattle in November

Abstract

"What is a gene?" is an important philosophical question that has been asked over and over. This paper approaches this question by understanding it as the individuation problem of genes, because it implies the problem of identifying genes and identifying a gene presupposes individuating the gene. I argue that there are at least two levels of the individuation of genes. The transgenic technique can individuate "a gene" as an individual while the technique of gene mapping in classical genetics can only individuate "a gene" as a type or a kind. The two levels of individuation involve different techniques, different objects that are individuated, and different references of the term "gene". Based on the two levels of individuation, I discuss important philosophical implications including the relationship between individuality and individuation and that between individuals and kinds in experimental contexts. I also suggest a new gene conception, calling it "the transgenic conception of the gene."

Keywords: gene concept, individuality, individuation, experiment, classical genetics, transgenic technique

1. Introduction: what is a gene and why individuation matters

"What is a gene?" and its related questions have been asked over and over by philosophers, historians, and scientists of biology (Beurton, Falk, and Rheinberger 2000; Carlson 1991; Falk 1986, 2010; Gerstein et al. 2007; Griffths and Stotz 2006, 2013; Kitcher 1982, 1992; Pearson 2006; Stotz and Griffiths 2004; Snyder and Gerstein 2003; Waters 1994, 2007). Those questions are frequently embedded in discussions about the definition of the term "gene" and the gene concept. As a consequence, the phrase "a gene" in this question usually refers to a type of gene. However, should we use "a gene" to refer to an individual gene, i.e., a gene token? Could it in fact be this?

The question "what is a gene" explicitly implies the problem of identifying genes, and identifying a gene presupposes individuating the gene. In what ways are genes individuated and how do scientists individuate them? I call this *the individuation problem of genes*. This paper shall approach the problem from three different but related perspectives.

From the epistemic perspective, a concept of the gene provides at least a working definition, which by nature is a hypothesis, for scientific research. Any hypothesis of the gene may be in error and may be confirmed only by experimentally individuating particular tokens of some gene. From the semantic perspective, according to a Fregean philosophy of language, the concept of reference usually serves for proper names that refer to individuals or particulars. We may extend the concept of reference to general terms (e. g., "humankind" or "gene kind") for the case in which some token of a kind is presented, and so we use a general term to refer to the kind. This means that at least some token of a kind has to be individuated. This semantic perspective presupposes an ontological perspective: the existence of a kind should be presented or demonstrated by the existence of at least a token of the kind. In the case of the gene, the ontological requirement means that we have to individuate a token of some gene kind. All three perspectives indicate the key status of individuation for answering the question of what a gene is.

According to the literature of analytic metaphysics, "individuation" is understood in a metaphysical and an epistemic sense. In the epistemic sense, someone individuating an object "is to 'single out' that object as a distinct object of perception, thought, or linguistic reference." (Lowe 2005: 75) This epistemic sense presupposes the metaphysical sense, in which what 'individuates' an object "is whatever it is that makes it the single object that it is – whatever it is that makes it one object, distinct from others, and the very object that it is as opposed to any other thing." (Lowe 2005: 75) Bueno, Chen, and Fagan (2018) add a practical sense to the term, interpreting

"individuation" as a practical process through which an individual is produced. They characterize the relation between "individuation" and "individuals" as when "an individual emerges from a process of individuation in the metaphysical sense. Epistemic and practical individuation, then, are processes that aim to uncover stages of that metaphysical process." (Beuno, Chen, and Fagan 2018) The approach to the individuation of genes I adopt herein follows their characterization, especially by focusing on the process of epistemic and practical individuation. Reversely, the case I am investigating in this paper offer an illustration for the new sense of individuation.

Although philosophers have investigated concepts of the gene and its change by examining many cases in scientific practices, they have seldom considered the role that the transgenic technique developed in biotechnology may play in philosophical discussions. This paper explores experimental individuation of genes from the direction of that technique, considering the possibility that a gene is individuated as an individual in the relevant contexts.

This paper thus addresses two central questions: (Q1) In what sense, can we reasonably say that classical geneticists have individuated a gene? (Q2) Are there experiments that can individuate a gene as an individual? Some new questions such as the relationship between individuality and individuation will be derived from the answer to the two questions. This paper is thus structured in the following way.

In the second section, I review the literature about the concepts and references of genes. Section 3 argues that the answer to Q1 is that the geneticists individuate a gene as a type, because they used the chromosomal location technique. Section 4 argues that the answer to Q2 is the experiments that use the transgenic technique. The two answers indicate two different kinds of individuation: individuation of a type and individuation of an individual. This raises a new question about whether or not "individuation of a type" is a consistent phrase. In order to respond to this, section 5 discusses in what sense we individuate a type and compare between two kinds of individuation defined by two different kinds of experiments and techniques: the chromosomal locution of genes and the transgenic experiment. My argument thus involves the relationship between kind and individual in the context of experimentation. Given the new question, Section 6 argues that transgenic experiments can demonstrate a gene type by individuating its tokens, while gene mapping experiments in classical genetics only individuate gene types. Thus, a new gene conception, calling it "the transgenic conception of the gene," can be proposed. I further discuss the relationship among the classical gene concept, the molecular gene concept, and the transgenic conception. In the seventh section, I defend the thesis that practices of individuation in scientific investigations are prior to characteristics of individuality identified by traditionally metaphysical speculations.

2. Concepts and references of the gene

The rapid change of the gene concept has produced a large multitude of gene concepts that have bewildered scientists (Gerstein et. al. 2007; Pearson 2006; Stotz and Griffiths 2004). The confused situation has attracted many philosophers and scientists to provide clarifying analyses. Although scientists as well as philosophers have made endeavors to overcome the predicament, they are motivated differently. Scientists believe that they need a unified concept to help them conduct research and to communicate with each other, because, as developmental geneticist William Gelbert says, "it sometimes [is] very difficult to tell what someone means when they talk about genes because we don't share the same definition" (Pearson 2006: 401). Thus, most scientists seek to redefine the "gene" and tend to adopt a single preferred perspective on the gene concept, although they are well aware with the plurality of gene definitions (Wain et. al. 2002; Gerstein et. al. 2007).

Philosophers at different times have been interested in clarifying concepts of the gene and in investigating the patterns of associated conceptual change. In contrast to actual definitions used by working scientists, they often consider more abstract concepts of the gene that can guide several different definitions in the context of scientific research. Consequently, they conclude that it is almost impossible to find a unified concept of the gene, and hence they take different stances to respond to this situation (rf. Waters 2007). Some are gene skeptics (e.g., Kitcher 1992). Some take a dualistic position, such as Moss (2003), who distinguishes between Gene-P and Gene-D based on the fields in that gene concepts are applied. Some are pluralists, such as Griffiths and Stotz (2006, 2013), who differentiate between three senses of the gene: the instrumental gene, the nominal molecular gene, and the postgenomic molecular gene. Still others are both pluralists and pragmatists. Waters (2018) emphasizes that scientists do and should apply different gene concepts under various investigative contexts.

With some exceptions, few philosophers explore the reference problem of the term "gene". Although Fregean semantics holds that the sense/concept or intension of a name determines its reference or extension, the matter about how a sense determines the reference is not easily seen from the scientific context. The determination of a theoretical term's reference usually involves experimental procedures and techniques that should be investigated and analyzed. Weber (2005, ch.7) does impressive work by providing several reference-determining descriptions of the term "gene" in the history of genetics. Based on those descriptions and the analysis of *Drosophila* genetic practices, he suggests that the pattern of referential change for "gene" is a kind of

freely floating reference. He also argues that different gene concepts refer to *different* natural kinds, which are overlapping but not coextensive.[1] According to Weber, reference for "gene" is fixed in the following manner for classical and molecular genes.

> Reference of [classical] "gene" (2): Whatever (a) is located on a chromosome, (b) segregates according to Mendel's first law, (c) assorts independent of other genes according to Mendel's second law if these other genes are located on a different chromosome, (d) recombines by crossing-over, (e) complements alleles of other genes, and (f) undergoes mutations that cause phenotypic differences. (Weber 2005: 210)

> Reference of [molecular] "gene" (5): The class of DNA sequences that determine the linear sequence of amino acids in a protein. (Weber 2005: 212)

Both classical and molecular gene concepts do refer to natural objects, because, as Weber notes (2005: 210-211), some *tokens* satisfying the reference-determining descriptions are experimentally presented when using the concepts with the intention of referring to sets of entities in historical occasions. However, one should note that the experimented tokens in classical genetics seems to be only some organisms with specific phenotypes (say, fruit flies or other kinds of organisms) while the experimented tokens in molecular biology may be some DNA segments. This difference raises interesting problem: what tokens are individuated in different contexts of experiments?

Before moving to the next section, I want to clarify that the individuation problem of gene concept's tokens is not the issue of gene individuality as raised by Rosenberg (2006: 121-133).[2] He defends the gene individuality thesis in parallel to the species individuality thesis, but Reydon (2009) objects to his argument and defends the gene as a natural kind. This paper aims to discuss how a gene kind and its tokens are individuated rather than whether or not an allele such as *Hbf* (the human fetal hemoglobin gene) is an individual.

3. Chromosomal location of a gene

[1] Baetu (2011: 411) argues that "the referents of classical and molecular gene concepts are coextensive to a higher degree than admitted by Waters and Weber…" However, Baetu builds his argument in terms of Benzer's work on phage. In my view, he does not successfully refute Waters' and Weber's arguments, because the referential change occurred within the classical gene concept, as Weber cogently argues.

[2] Rosenberg uses "natural selection and the individuation of genes" as the title of the section in which he discusses the gene individuality thesis.

Weber's argument indicates that we may and should consider the reference of the classical gene concept independently of the molecular gene concept and others. Weber's reference-determining description of "gene" (2) indicates that the chromosomal location (or mapping) of genes plays a key role in determining referents. However, the question "what tokens are individuated and thus referred to?" does not be answered.

Classical geneticists in the early 20th century located and labeled some specific classical genes on some specific chromosomes. The earliest genetic map (see Figure 1) of *Drosophila melanogaster* (fruit fly) was depicted in 1915. Figure 1 shows that the gene (allele) pair of *Drosophila's* grey body and (mutant) yellow body is located at the first locus on the first chromosome. The second gene pair of red eyes and (mutant) white eyes is located below the grey body gene. The other genes are located below the first two in order. However, every gene is differently distant from the first gene and thus occupies *a single locus* without overlapping. Accordingly, are we able to say that the location of a gene individuates the gene? Before answering this question, it is necessary to discuss how classical geneticists locate a gene on a chromosome. In other words, what technique is used in the process of locating genes?

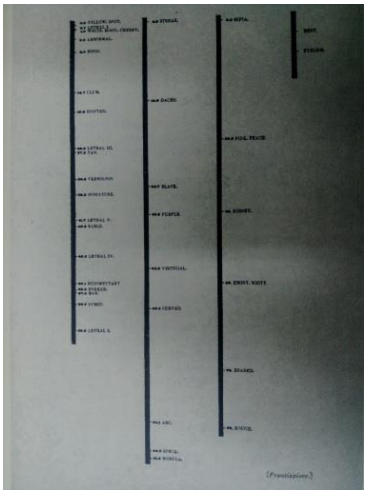


Fig. 1. Genetic map of *Drosophila* in 1915. Reproduced from Morgan, T. H. et. al. (1915).

Chromosomal location or mapping of genes is a well-known story (Darden 1991, Waters 2004, Weber 2005, 2006; Falk 2009). For the purpose of this paper, I introduce a very brief version. In the 1910s, Thomas Hunt Morgan's team developed a

technique to map the linear relations among factors (genes) in linkage groups, using Mendelian breeding data. Morgan and his team discovered that a pair of chromosomes may cross over with each other partially during the period of meiosis. Crossing over produces a specific ratio of the linked traits. Morgan believed that "the percentage of crossing over is an expression of the 'distance' of the factors from each other." (Morgan et.al. 1915: 61) Sturtevant then used percentages of linked characters that exhibited crossing over to calculate the relative positions of the factors to each other. This is the kernel technique for constructing genetic maps. By using genetic maps, Morgan's team determined the loci of many genes on the four chromosomes of *Drosophila*. Given the genetic maps, the classical geneticists assume that no other genes are located at the same position of a chromosome.[3] As a consequence, the single location of a gene actually indicates the individuality of genes.

Genetic maps by nature are diagrammatic models for the actual loci of genes in chromosomes. They are inferences from the statistical data of breeding experiments. Models represent the general. When we say that the location of a gene in a genetic map represents the locus of a classical gene on a chromosome, we really mean that it represents the locus of a type of classical gene on an identical type of chromosome in a cell within a kind of organism. Of course, this implies that a token of a type of classical gene on a token of a type of chromosome can be cognitively identified and discerned, because we can distinguish it from the tokens of the other genes. As a result, we can also count genes within cells. The located genes thus satisfy the two traditional characteristics of individuality: distinguishability and countability.[4]

If all chromosomes were stick-shaped substances of uniform material without complicated structure, then the chromosomal location of classical genes would be able to genuinely individuate them. According to molecular biology, however, chromosomes are a long chain of double helix DNA molecules that curl themselves up in twisted shapes. In such a case, we cannot delineate a located classical gene or depict its contour or boundary, because the chromosomal locus at which the gene is located includes a twisted part of the long DNA molecule. Even by invoking the knowledge from molecular biology, one would still be puzzled by the problem of defining the molecular gene.

4. Individuating molecular genes as individuals

Ever since the era of molecular biology, the continuously accumulating knowledge of genetics has not solved the individuation problem of genes. Instead, it

[3] Of course, a full story is more complicated. For the simplifying purpose, I skip the relevant discussion about gene mutation.

[4] The implications of using these criteria will be discussed in the sixth section.

has brought more troubles about the definition of the gene concept. Is a gene "a sequence of DNA for encoding and producing a polypeptide"? Should we include the start and stop codons (i. e., the regulation problem)? Should we count those introns deleted during the process of transcription into the investigated gene (i.e., the splicing problem)? The difficulty in defining the molecular gene concept directly contributes to the impediment of individuating a gene.

Many gene sequencing projects have been conducted during the genomic era. Scientists do not identify a DNA sequence as a gene and discern the gene from others by using gene sequencing *per se*, because it offers only syntactical orders of genetic codes. Gene annotation, which is used to infer what those annotated sequences do, has been developed to offer *senses* or *intensions* for them. However, the impediment of discerning genes remains, because the definition of the gene is still vague and confusing (rf. Baetu 2012; Gerstein et. al. 2007; Griffiths and Stotz 2013, ch. 4). In fact, gene annotation is based on several assumptions, by which scientists infer that a few sequences may be genes that contribute to phenotypes or functions. Those assumptions need to be confirmed by experimental investigations. Many techniques such as directed deletion, point mutation making, gene silencing, and transgenesis in reverse genetics have been developed to determine what a gene is and what it does (Gilchrist and Haughn 2010).

I argue that the transgenic technique is a very definite and powerful way to individuate a gene. It can even individuate molecular genes as individuals without a clear boundary of a gene or a clear definition of the gene, although the technique is limited.[5] How does the transgenic technique do this? What conditions of individuality allow the technique to individuate a gene as an individual?

Chen (2016) proposes a conception of experimental individuality with three attendant criteria (separability, manipulability, and maintainability of structural unity) and argues that the first experiment of bacteria transformation individuated an antibiotic resistance gene by satisfying the three criteria.[6] Below I reiterate this story in brief.

Stanley Cohen and Herbert Boyer combined DNA of *Escherichia coli* (*E. coli*) in 1973 and 1974 by transferring two different DNA segments encoding proteins for ampicillin and tetracycline resistance into *E. coli*, thereby realizing the transformation of this bacterium (Cohen et. al. 1973; Chang and Cohen 1974). Both DNA segments are called an "antibiotic resistance gene." Cohen and Boyer used small circular

[5] The technique cannot be applied in many occasions because of technological difficulties. It should not be applied to humankind due to ethics consideration. In addition, many gene-modification organisms produced by using the technique may involve ethical issues.

[6] Chen (2016) uses the creation of Bose-Einstein condensates in physical experiments as the other example. Chen's intent is to argue that biological entities and physical entities in laboratories share the same criteria of experimental individuality.

plasmids (extrachromosomal pieces of DNA) as vectors to transfer a foreign DNA segment into a bacterial cell. The plasmids were made by cutting out a (supposed) antibiotic resistance gene from other bacteria with the restriction enzyme *EcoRI*, linking the segment into a plasmid by using another enzyme, DNA ligase. The scientists then transferred the plasmid into an *E. coli* cell without the ability to resist antibiotics. The result, a modified *E. coli* cell, was able to resist antibiotics and contained the antibiotic resistance gene. In that experiment, the antibiotic gene was separated from its original bacteria and then was manipulated (i.e., linked and transferred). Its structural unity was not broken down, hence allowing it to be expressed in the other kind of bacteria. Scientists thus identify it as a gene, an individual biological entity, because the separated, manipulated, and maintained antibiotic gene was naturally separable, manipulable, and maintainable. The photos in Figure 2 show that scientists worked with a single DNA segment, as indicated by (b) in [A] and [B].

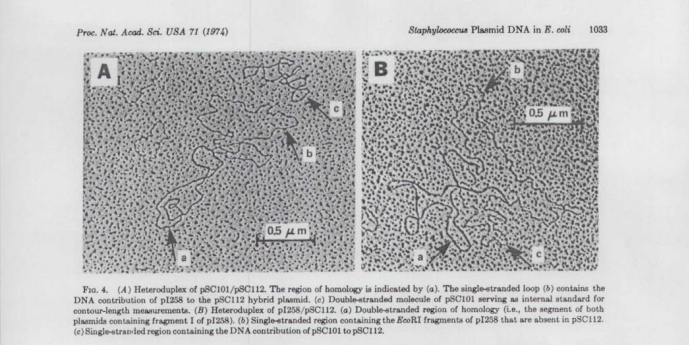


Fig. 2. Two pictures of plasmids in bacterial transformation. Reproduced from Chang and Cohen (1974).

I next interpret the performance of the technique used in transgenic experiments as the general process of individuating transgenes. The process has five stages.

(1) Use restriction enzymes to cleave specific segments from recognition sites of long DNA chains. A specific restriction enzyme can cut away a specific DNA segment at a specific site.

(2) Link the cleaved segment of DNA to a plasmid vector by using DNA ligase. The vector is a circular DNA that may come from a wild type of virus.

(3) Incorporate the DNA segment in the vector into the genome of another organism by injecting the plasmid vector to a cell of the target organism. Of course,

they may fail when the intended feature is not expressed.

(4) Make copies of DNA segments by cloning the cell containing the transferred segment of DNA. The aim of DNA cloning is to copy a segment of interest (or a gene) from an organism and produce many copies.

(5) Observe the expression of the novel feature that the target organism does not typically have. If a DNA segment cut from an original organism is successfully pasted into a cell of a target organism and the target organism expresses the intended feature that the original organism has, then one concludes that the segment is a gene.

The first stage corresponds to the separation condition, the second, the third, and the fourth stages to the manipulation condition, and the fifth stage to the maintenance condition. Accordingly, one can easily see that those cut, linked, transferred, pasted, and copied genes are particulars – individuals, because they satisfy the three criteria of experimental individuality that indicates their *singularity* and *particularity*. In other words, a single segment of DNA maintains its structural unity when being separated and manipulated. This is so, because cutting a gene from an original organism is in fact separating it from its environment and because transferring, pasting and copying a gene is manipulating it. If the gene does express the intended feature in a target organism, then this condition indicates that the unity of its chemical and informational structure has been maintained.

5. Two kinds of individuation of genes

The previous discussion indicates that two different objects have been individuated in different experimental and theoretical contexts. In the context of classical genetics, scientists used breeding experiments and theoretical inferences to locate a gene at some locus on a chromosome. They would individuate genes as types if they assume that no other genes could coexist at the same locus. If one interprets the meaning of "individuation" as "only individuals can be individuated," then the phrase "individuating genes as types" sounds unreasonable. Is it better to say "unitization of genes" rather than "individuation of genes"?

It is quite right to say classical geneticists *unitize* genes as types. In a sense, however, we may reasonably say that we individuate a gene as a type, because the type has tokens or members that are distinguishable and countable individuals. Classical geneticists suppose that all types of genes have corpuscular members, i.e., substantive individuals. In such a sense, talking of "individuating genes as types" is reasonable. If no distinguishable and countable members or samples of a kind can be identified, then the kind cannot be individuated. In other words, we cannot individuate

such a kind as water or air that is expressed by "mass" nouns at the macroscopic level, although we can individuate a sample of water by using a container or individuate a water molecule by specific technique at the molecular level. For the cases of experiments using the transgenic technique, molecular biologists physically individuate *singular and particular* gene tokens. Thus, we claim that scientists experimentally individuate genes as individuals in such a context.

In consequence, two different sets of criteria for individuality are presupposed. Experiments using the location technique have individuated a type whose tokens or members are countable individuals rather than matter referred to by mass nouns. In such experimental contexts, we emphasize distinguishability and countability as the indexing features of individuals. Experiments using the transgenic technique individuate singular and particular individuals – gene tokens. For these experimental contexts, we emphasize singularity and particularity of individuals in contrast to universality of types or kinds. We assure the particularity and singularity of the individuals through the realization of experimental individuality, namely, the joint realization of separability, manipulability, and maintainability of structural unity. At this point, more philosophical implications will be discussed in next section.

The two individuated targets indicate two different referential levels of the term "gene" in the literature. As we have seen, when many philosophers and scientists ask "what is a gene," they really refer to a type of gene in conjunction with discussing the gene concept or the definition of "gene." Similarly, in some contexts of scientific investigation, scientists use "a gene" to refer to a type of gene as the phrase "chromosomal location of a gene". In the context of transgenic experiments, however, "a gene" is used to refer to a genuine individual – a single and particular gene token, because scientists have worked with particular objects that maintain their structural unity when being separated and manipulated in the process of experimenting.

The two referential levels indicate two different kinds or levels of experimental individuation, which are realized by two different techniques: the chromosomal location technique and the transgenic technique. Although the two techniques aim to the same target (i.e., genes or types of genes), they physically experiment and manipulate different objects. Experiments using the chromosomal location technique indirectly identify loci of genes by manipulating organisms that contain chromosomes with genes in breeding, while experiments using the transgenic technique directly manipulate DNA segments. Therefore, classical geneticists can only cognitively discern gene types by identifying their loci without practically interacting with gene tokens; they really practically interact with organismal individuals that contain different types of genes. Reversely, molecular biologists can practically interact with gene tokens and then cognitively infer out the existence of a gene type.

6. Gene concepts and individuation

One may still wonder: Can the location technique individuate a singular and particular gene in the sense of individuating entities as individuals? The answer is obviously negative, because that technique cannot separate and manipulate a gene token and maintain its structural unity. On the contrary, one may ask: Can the transgenic technique individuate a type of gene? Here the answer is less clear. In the sense that scientists suppose that a token of a gene has been physically individuated in transgenic experiments, we are allowed to say that the technique also individuates a type of gene. However, scientists are not fully sure that the transgenic technique on a posited gene can be always successfully applied to another individual of the same organism. In fact, the probability of failure is quite high. Unless the experimental individuation of particular tokens can be performed repeatedly and stably, then one can say that the gene tokens indicate a general type of gene and that the type has been identified. However, the object individuated by the technique is not a type of gene, because the technique always requires manipulating particular segments of DNA -- gene tokens. If a kind of transgenic experiment with a specific transgene has been stably repeated, then a type of gene has been discovered by experimentally individuating its tokens in performing such an experiment.

Since transgenic experiments may be successfully and stably performed by using different transgenes, one can extract a special conception of the gene that is characterized by the transgenic technique. I call this "the transgenic conception of the gene," in which *a gene is a transferrable DNA sequence which is able to express a phenotype/function on another kind of organisms*. Of course, this does not imply that those technically untransferrable DNA sequences are not genes, given the fact that the number of transgenes is relatively few to the number of genes located at chromosomes. This is so because scientists do not always find the precise site of a gene (type) and available restriction enzymes to cut the DNA segment of the gene. Thus, the extension of the transgenic conception of the gene is not equivalent to that of the classical gene concept. Due to the limited number of transgenes, the transgenic conception is not yet co-extensional with the molecular gene concept. To be precise, the extension of the former is included within the extension of the latter, because all transgenes are molecular genes but not all molecular genes can be transplanted. In addition, the intension of the transgenic conception is implied in the intension of the molecular gene concept, because the technique was developed from molecular biology. As a consequence, the transgenic conception can be viewed as a *sub-conception* of the molecular gene concept. Nevertheless, we have a conception

derived from scientific practices.

7. The priority of individuation to individuality

Bueno, Chen, and Fagan (2018) promote an approach by which investigating processes of individuation in scientific practices is prior to metaphysical speculation on criteria of individuality. This paper obviously follows the approach. However, this does not mean that we do not need any criterion of individuality in identifying any individual in scientific practices. Rather, criteria of individuality are implied in or extracted from procedures of scientific practices, as the three conditions of experimental individuality are extracted from experimental practices (Chen 2016). Criteria of individuality based on scientific practices may or may not conflict with criteria from metaphysical theories. Considering the relationship between practical criteria and speculative criteria will help us understand practical individuation more deeply.

The metaphysical tradition has identified at least six characteristics or indexing features of individuality in general: particularity, distinguishability, countability, delineability, unity, and persistence (Pradeu 2012: 228-229; Chen 2016: 351).[7] Recently, some philosophers argues that all biological entities are processes (Dupré 2018, Nicholson and Dupré 2018, Pemberton 2018), so I would like to add processuality to the list. Indeed, I believe that all biological individuals pass through a life, i.e., a process (see also Chen 2018), therefore, processuality is a central characteristic of biological individuality. Those characteristics, originally come from metaphysical speculation, can singly, jointly, or collectively serve as epistemic criteria of individuality.

In the context of scientific practices, they are the outcomes from rather than preconditions for the realization of individuation. For example, individuating genes as individuals in the context of transgenic experiments indicates that the separated, manipulated, and maintained genes are particular and singular tokens. As the experimental individuation of gene tokens is realized, those tokens are also distinguishable, countable, unitary, persistent, and passing through a process, because particular and concrete individuals are being separated, manipulated, and maintained. The practices of separation and manipulation indicate epistemic particularity,

[7] Characteristics of individuality can serve as criteria of individuality and thus be involved in a theory of individuation. Bueno, Chen, and Fagan (2018) identify six theories of individuation in traditionally analytic metaphysics. A theory of individuation in the metaphysical sense involves not only "a theoretic construction of the nature of individuality and its attendant criteria," but also other metaphysical concepts such as "property, trope, universal, particular, substance, substratum, time, space, sort or kind." (p. 3) For my purpose, I will discuss only characteristics of individuality rather than any theory of individuation.

distinguishability, and countability. The practice of maintenance of structural unity indicates the unity, persistence, and processuality of the maintained gene token. However, all of the three practices would not indicate the delineation of a gene token, because the spatial boundary of the manipulated gene does not and cannot be delineated. Of course, this point does not mean that delineation is not a characteristic of individuality, but rather that it is not applicable to this case.

Individuating genes as types in classical genetics indicates that the individuated types of genes contain distinguishable and countable tokens, because the individuation is the location of a gene at a chromosome in a diagrammatic model. Supposing that the loci of different genes do not overlap, then the special locus of a gene is thus distinguishable from the locus of another gene. As a consequence, a gene token at a chromosome in a cell of a kind of organism is thus distinguishable from another token of the identical type of gene. All gene types located at chromosomes are countable. Supposing that every organism contains a token of a specific type of gene, then tokens of that gene type are countable. However, chromosomal location of genes does not indicate particular and singular gene tokens, because the individuated objects are only types of genes. As I have argued, the kind of individuation practice did not touch down the manipulation of individuals and remained in the cognitive level which focuses on gene types in general.

Although the concept of individuation can be reasonably applied to a kind whose members are individuals, all characteristics of individuality are not applicable. One cannot apply particularity, delineation, unity, and processuality to gene types, because a gene type is, in principle, universal, occupying multiple spaces, not cohesive, replicable, and non-processual. However, distinguishability and countability can be adequately applied to gene types, because one can distinguish one gene type from another gene type and count gene types when the chromosomal location is realized. In this case, thus, both distinguishability and countability cannot sufficiently demonstrate that the individuated objects are individuals. On the other hand, in the case of transgenic experiments, we can derive particularity, unity, and processuality from the three conditions of experimental individuation (separation, manipulation, and maintenance of structural unity). As a consequence, characteristics of individuality are derived from individuation; they are outcomes of practical individuation.

8. Conclusion

In this paper, I argue that there are at least two kinds of experimental individuation of genes. Scientists individuate genes as types in classical genetics and

individuate genes as tokens in transgenic experiments. Individuating a gene as a type or individuating a gene as an individual depends on the technique used in experimentation. I argue that characteristics of individuality identified in traditional metaphysics are not presupposed by individuation. Rather, they are outcomes or products derived from practical individuation in scientific experiments. I further argue that different kinds of experimental individuation presuppose different concepts of the gene: the classical gene concept and the transgenic conception of the gene. I argue that the transgenic conception can be viewed as a sub-conception of the molecular gene concept. An outstanding problem remains. Whether we can unify different concepts of the gene by integrating different experimental techniques, such as the chromosomal location technique, the technique of genetic sequencing, the techniques in reverse genetics, and the transgenic technique. Future analyses can approach this and other related questions in light of our new understanding of how classical geneticists individuated genes and the role experimental techniques play in identifying a gene as an individual.

Acknowledgment: I thank Alan Love, Ken Water, and Marcel Weber for their very valuable comments and suggestions. This paper has been revised according to their comments.

References

Baetu, Tudor M., 2011. “The referential convergence of gene concepts based on classical and molecular analysis,” *International Studies in the Philosophy of Science*, 24 (4): 411-427.

Baetu, Tudor M., 2012. “Genes after the human genome project.” *Studies in History and Philosophy of Biological and Biomedical Science*, 43: 191-201.

Beurton, P., R. Falk, and H.- J. Rheinberger, 2000. *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*. Cambridge, UK: Cambridge University Press.

Beuno, Otavio, Ruey-Lin Chen, and Melinda B. Fagan, 2018. “Individuation, process, and scientific practice.” In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 1-18. New York: Oxford University Press.

Carlson, E., 1991. “Defining the gene: an evolving concept.” *American Journal of Human Genetics*, 49: 475-487.

Chang, Annie C. Y. and Stanley N. Cohen, 1974. “Genome construction between bacterial species *in vrito*: Replication and expression of *Staphylococcus* plasmids

in *Escherichia coli*," *Proceedings of the National Academy of Science of USA*, 71(4): 1030-1034.

Chen, Ruey-Lin, 2016. "The experimental realization of individuality." In Alexandre Guay and Thomas Pradeu (eds.). *Individuals across the Sciences*, 348-370. New York: Oxford University Press.

Chen, Ruey-Lin, 2018. "Experimental Individuation: Creation and Presentation," In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, . New York: Oxford University Press.

Cohen, Stanley N. et. al., 1973. "Construction of biologically functional bacterial plasmids *in vitro*," *Proceedings of the National Academy of Science of USA*, 70(11): 3240-3244.

Darden, Lindley, 1991. *Theory Chang in Science: Strategies from Mendelian Genetics*. Oxford: Oxford University Press.

Dupré, John, 2018. "Processes, Organisms, Kinds, and Inevitability of Pluralism." In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 25-38. New York: Oxford University Press.

Falk, Raphael, 1986. "What is a gene?" *Studies in History and Philosophy of Science*, 17: 133-173.

Falk, Raphael, 2009. *Genetic Analysis: A History of Genetic Thinking*. Cambridge: Cambridge University Press.

Falk, Raphael, 2010. "What is a gene – revised" *Studies in History and Philosophy of Biological and Biomedical Science*, 41: 396-406.

Gerstein, Mark B. et. al. 2007. "What is a gene, post-ENCODE? History and updated definition." *Genome Research*, 17(6): 669-681.

Gilchrist, Erin and George Haughn, 2010. "Reverse genetics techniques: engineering loss and gain of gene function in plants," *Briefings in Functional Genomes*, 9(2): 103-110.

Griffths, Paul and Karola Stotz, 2006. "Genes in the postgenomic era," *Theoretical Medicine and Bioethics*, 27(6): 253-258.

Griffths, Paul and Karola Stotz, 2013. *Genetics and Philosophy: An Introduction*. Cambridge: Cambridge University Press.

Kitcher, P. S., 1982. "Genes." *British Journal for the Philosophy of Science*, 33: 337-359.

Kitcher, P. S., 1992. "Gene: current usages." In E. Keller and L Lloyd (eds.), *Keywords in Evolutionary Biology*. Cambridge, MA: Harvard University Press, pp. 128-131.

Lowe, E. Jonathan 2005. "Individuation," *The Oxford Handbook of Metaphysics*, ed. Michael J. Loux and Dean W. Zimmerman. Oxford: Oxford University Press.

Maienchein, J., 1992. “Gene: Historical perspectives.” In E. Keller and E. Lloyd (eds.). *Keywords in evolutionary biology*. Cambridge, MA: Harvard University Press, pp. 181-187.

Morgan, Thomas Hunt, et.al., 1915. *The Mechanism of Mendelian Heredity*. New York: Henry Holt and Company.

Moss, Lenny, 2003. *What Genes Can’t Do*. Cambridge, Mass.: The MIT Press.

Nicholson, Daniel J. and John Dupré, 2018. Everything flows: Towards Processual Philosophy of Biology.

Pearson, Helen, 2006. “What is a gene?” *Nature*, 441(25): 399-401.

Pemberton, John. 2018. “Individuating Processes,” In Otavio Beuno, Ruey-Lin Chen and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 39-62. New York: Oxford University Press.

Pradeu, Thomas, 2012. *The Limits of the Self: Immunology and Biological Identity*. Oxford: Oxford University Press.

Reydon, Thomas, 2009. “Gene Names as Proper Names of Individuals: An Assessment.” *British Journal for the Philosophy of Science*, 60(2): 409-432.

Rosenberg, Alexander, 2006. *Darwinian Reductionism*. Chicago: The University of Chicago Press.

Snyder, Michael and Mark Gerstein, 2003. “Defining genes in the genomics era.” *Science*, 300(5617): 258-260.

Stotz, Karola and Paul Griffths, 2004. “Genes: philosophical analyses put to the test.” *History and Philosophy of the Life Sciences*, 26: 5-28.

Wain, H. M., et. al. 2002. “Guidelines for human genome nomenclature,” *Genomics*, 79: 464-470.

Waters, Kenneth C., 1994. “Genes made molecular,” *Philosophy of Science*, 61: 163–185.

Waters, Kenneth C., 2004. “What was classical genetics?” *Studies in History and Philosophy of Science*, 35 (4): 783-809.

Waters, Kenneth C., 2007. “Molecular genetics,” *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/molecular-genetics/

Waters, Kenneth C., 2018. “Don’t Ask ‘What is an individual?’” In Otavio Beuno, Ruey-Lin Chen, and Melinda B. Fagan (eds). *Individuation, Process, and Scientific Practice*, 91-113. New York: Oxford University Press. (In press)

Weber, Marcel, 2005. *Philosophy of Experimental Biology*. Cambridge, UK: Cambridge University Press.

Weber, Marcel, 2006. “Representing genes: Classical mapping techniques and the growth of genetic knowledge,” *Studies in History and Philosophy of Biological and Biomedical Science*, 29: 295-315.

The Verdict's Out:

Against the Internal View of the Gauge/Gravity Duality

4993 words

Abstract

The gauge/gravity duality and its relation to the possible emergence (in some sense) of gravity from quantum physics has been much discussed. Recently, however, Sebastian De Haro (2017) has argued that the very notion of a duality precludes emergence, given what he calls the internal view of dualities, on which the dual theories are physically equivalent. However, I argue that De Haro's argument for the internal view is not convincing, and we do not have good reasons to adopt it. In turn, I propose we adopt the external view, on which dual theories are not physically equivalent, instead.

1 Introduction

The gauge/gravity duality has generated much discussion about whether space-time geometry or gravity emerges (in some sense) from quantum physics.[1] Recently, however, De Haro [2017] has argued that the very notion of a duality *precludes* the possibility of emergence given what he calls the *internal view* of dualities, on which dual theories are physically equivalent. In turn, this claim impinges upon the broader debate about whether we can make claims about emergence given a duality. After all, since the internal view of dualities is supposed to *rule out* emergence, any such debate is rendered moot once we adopt the internal view. My goal here, though, is to argue that De Haro's argument for the internal view is not convincing. Instead, I propose we adopt the *external view* of dualities, on which dual theories are *not* physically equivalent.

First, I introduce Fraser's [2017] three-pronged distinction of predictive, formal and physical equivalences, characterizing dualities in terms of this distinction (§2.1). I then make things more concrete by briefly considering the gauge/gravity duality via the Ryu-Takayanagi conjecture from the **AdS/CFT** (anti-de Sitter space/conformal field theory) correspondence (§2.2).

Next, I introduce De Haro's interpretive fork between the internal and external views of dualities (§3). I illustrate how the internal view is supposed to preclude emergence, but criticize De Haro's argument for the internal view – that it is meaningless to hold the external view given 'some form of' structural realism and how the two theories are

[1] One prominent physicist who is a proponent of emergent space-time is Seiberg 2007, while philosophers like Rickles 2011/2017, Teh 2013, and Crowther 2014 have all tackled the topic.

'totalizing' in some way – by showing how it does not work without further assumptions (§4). In turn, given the interpretive fork, I propose we adopt the external view instead. In concluding remarks, I briefly discuss this result in relation to the broader debate about emergence within the gauge/gravity duality.

2 Gauge/Gravity through AdS/CFT

2.1 Duality

Fraser [2017] takes two theories related by a duality to have two features: (i) they agree on the transition amplitudes and mass spectra, and (ii) there is a 'translation manual' that allows us to transform a description given by one theory to a description given by another theory. We may explicate (i) and (ii) by first considering distinct sorts of 'equivalence' proposed by Fraser [2017, 35]:

- *Predictive equivalence*: "there is a map from T_1 to T_2 that preserves the values of all expectation values deemed to have empirical significance by T_1 and that preserves the mass spectra, and vice versa."
- *Formal equivalence*: "there is a translation manual from T_1 to T_2 which maps all quantum states and quantum observables deemed to have physical significance by T_1 into quantities in T_2 and respects predictive equivalence, and vice versa."
- *Physical equivalence*: "there is a map from T_1 to T_2 that maps each physically significant quantity in T_1 to a quantity in T_2 with the same physical interpretation and respects both formal and predictive equivalence, and vice versa."

Given our characterization of a duality as (i) and (ii), we may quite naturally say that two theories are dual to one another when they are *predictively* and *formally* equivalent. Furthermore, supposing that this three-pronged distinction exhausts the possible equivalences relevant to physics, we might also say that two theories satisfying (i)-(iii) are also *fully*, or *theoretically*, equivalent.

Here it would be germane to differentiate two distinct sorts of structures in a duality. Given predictive and formal equivalence, the isomorphism holding between physical and empirical quantities of the dual theories suggests a structure, which may be called the *empirical core* of the duality. However, as Teh [2013, 301] also notes, despite the empirical core, "duality is precisely an equivalence between two theories that describe (in general) different physical structures, i.e. theories with non-isomorphic models." In other words, while there is an empirical core, by which physical and empirical quantities are mapped onto one another, these quantities are generally related to other quantities in a quite different manner on each side, viz. there is 'excess structure' exogenous to the empirical core. Without further argument, we are not entitled to 'discard' this 'excess structure', which also means that predictive and formal equivalence (characterizing the empirical core) does not automatically entail physical, and hence full, equivalence.

Given Fraser's framework, I will briefly introduce the gauge/gravity duality more concrete by briefly examining the example of **AdS/CFT** correspondence.

2.2 The AdS/CFT Correspondence

The *gauge/gravity duality*, or *holographic principle*, postulates a duality between a suitably chosen N-dimensional gauge quantum field theory (QFT) that does not describe

gravity, and a quantum theory of gravity in $(N+1)$-dimensional space-time (the 'bulk') with an N-dimensional 'boundary', on which the gauge theory is defined. Hence the slogan: gauge on the boundary, gravity in the bulk.

The **AdS/CFT** correspondence is a specific case of the gauge/gravity duality. On the one hand, '**AdS**' stands for anti-de Sitter space-time - a maximally symmetric solution to the Einstein equations with a constant negative curvature and a negative cosmological constant. More accurately, though, the '**AdS**' in **AdS/CFT** correspondence should be taken to refer to a string theory of quantum gravity defined *on* a 5-dimensional **AdS**. '**CFT**', on the other hand, refers to a quantum field theory with scale (or conformal) invariance defined on the 4-dimensional boundary of the **AdS**. The **AdS**-side theory is defined in the 'bulk', and the **CFT**-side theory is defined on the 'boundary' of the **AdS** space-time.

The **AdS/CFT** correspondence, then, refers to a postulated duality between the two theories, satisfying (i) and (ii) from §2.1. (i) is satisfied given the postulate that bulk fields propagating in the bulk are coupled to operators in the boundary **CFT**. Hence, the **AdS** theory of gravity will predict exactly the 'same physics', viz. transition amplitudes, expectation values and so on, as the **CFT** theory without gravity.

Beyond empirical, i.e. measurable, quantities, physically significant quantities of **AdS/CFT** must also relate to one another since it is a duality. In other words, (ii) is supposed to hold simply as a core postulate. This is not to say that (ii) is completely unfounded: in particular, we have evidence suggesting that at least *some* physical quantities of dual theories are related to one another in surprising ways, which in turn supports the claim that (ii) holds. Here I will focus on one such relation, the Ryu-Takayanagi conjecture.

The Ryu-Takayanagi conjecture postulates that the entanglement entropy of two regions on the boundary is related to the surface area within the bulk:[2]

$$(\mathbf{RT}):\ S_A = \frac{Area(\widetilde{A})}{4G_N}$$

RT tells us that the entanglement entropy of a region on the boundary of the **AdS**, S_A, viz. the von Neumann entropy[3] in the **CFT**, is directly proportionate (by 4 times the Newtonian gravitational constant) to the area of the boundary surface $\widetilde{A}$ bisecting the bulk, dividing the two entangled regions on the boundary. Below, *Fig. 1.* shows a simplified diagram for visualizing **RT**.

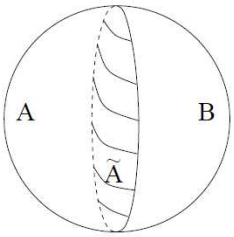


Fig. 1. The area $\widetilde{A}$ bisects the bulk space-time into two, and on the boundaries of the two parts we define the regions A and B. The Ryu-Takayanagi formula tells us that given a change in S_A we get a change in the size of $\widetilde{A}$ by the proportion of $\frac{1}{4G_N}$. [Figure taken from Van Raamsdonk 2010]

RT paints an interesting picture for emergence of space-time geometry from quantum theory: the area of a space-time itself is closely related to quantum entanglement entropy in a surprising way. An increase in the entanglement entropy between two

[2]See Ryu & Takayanagi 2006 for technical details.

[3]The von Neumann entropy is given by $S_A = -Tr(\rho_A log \rho_A)$. The reduced density matrix describing the region A, ρ_A, is obtained from tracing over the B-components of the combined density matrix of A and the entangled region B, ρ_{AB}: $\rho_A = Tr_B(\rho_{AB})$.

regions of a field described by **CFT** leads to a proportionately increasing boundary area of the bulk, and hence a geometric (or gravitational) phenomenon is described in terms of a quantum phenomenon.[4]

Given relations like **RT**, we can also see more clearly how **AdS/CFT** is supposed to satisfy (ii): physically significant quantities, such as 'area' of space-time in the bulk and 'entanglement entropy' between two regions on the boundary, are mapped to one another via suitable equations. Hence, **AdS/CFT** is a special case of the gauge/gravity duality: a theory of quantum gravity on a $(N+1)$-dimensional **AdS** space-time is dual to a **CFT** defined on its N-dimensional boundary.

With the gauge/gravity duality made concrete, let us turn to the interpretive task.

3 The Internal View

Dieks et al. [2015] and De Haro [2017] proposes an interpretive fork for dualities: we can either adopt an internal or external view. De Haro describes the *internal view* as such:

> if the meaning of the symbols is not fixed beforehand, then the two theories, related by the duality, can describe the same physical quantities. [...] we have two formulations of one theory, not two theories. [De Haro 2017, 116]

On the contrary, the *external view* holds that:

> the interpretative apparatus for the entire theory is fixed on each side. [...] On this interpretation there is only a formal/theoretical, but no empirical, equivalence between the two theories, as they clearly use different physical

[4]See Van Raamsdonk 2010 for an excellent summary of this picture.

> quantities; only one of them can adequately describe the relevant empirical observations.

Is De Haro's characterization of the external view adequate? The fact that there is no 'empirical' equivalence (what Fraser calls physical equivalence) between two theories does not entail that at most one of them can adequately describe the relevant empirical observations, where one description is 'correct' and the other 'wrong', nor does it entail mutually exclusive physics where only one theory can be correct at any one time. To assume so seems to rule out, by fiat, the possibility of emergence, since emergence relies on *both* theories being in a way adequately descriptive of the world (except one is more 'fundamental' than the other). Hence, taking in account Fraser's framework, I re-characterize the external view as such: it is simply the claim that the two dual theories are *physically non-equivalent* i.e. have distinct physical interpretations, despite formal and predictive equivalence.

Given the interpretive fork, if we are led to forsake the internal view, then we are motivated to accept the external view instead. As such, my strategy here is to show that we should forsake the internal view, and in turn accept the external view instead.

To better understand what the internal view is claiming, I break it down into three constituent claims.

The first claim is that of *theoretical equivalence*: under the gauge/gravity duality, the two theories (e.g. **AdS** and **CFT**) are taken to be simply different formulations of a single theory, describing the same physical quantities despite their obvious differences. As Dieks et al. puts it, 'the two theories collapse into one' [2015, 209-210]. In light of Fraser's framework described in §2.3, this claim means that the gauge/gravity duality, on

the internal view, involves the conjunction of predictive, formal and physical equivalences. In other words, beyond a one-to-one mapping (a 'translation manual') of relevant physical quantities and the sharing of all transition amplitudes, mass spectra and other observable predictions, the internal view claims that the two theories also have the *same physical interpretation*. However, as Fraser [2017, 35] notes, "predictive equivalence does not entail formal equivalence, and formal equivalence does not entail physical equivalence." Formal and predictive equivalence cannot entail physical equivalence on their own.

The internal view's claim of theoretical equivalence, then, must require an additional claim of *physical equivalence*, in addition to formal and predictive equivalence: the dual theories are taken to be physically equivalent, and hence have the same physical interpretation. As per §2.1, this would indeed entail theoretical equivalence.

Physical equivalence is in turn justified by a third claim, that the two theories in a duality should be left *uninterpreted*. As De Haro claims above, assume 'the meaning of the symbols is not fixed beforehand'. Then, given formal and predictive equivalence, we have an isomorphism between the dual theories' (now-uninterpreted) 'physical quantities' and numerical predictions, viz. an uninterpreted empirical core. Ignoring the 'excess structure' exogenous to the empirical core, we can then take the empirical core to be representing a single uninterpreted theory, where the now-uninterpreted 'quantities' of each dual theory now refer to the 'places' or 'nodes' of the empirical core's structure. As Dieks et al. (2015) puts it,

> A in one theory will denote exactly the same physical quantity as B [...] if these quantities occupy structurally identical nodes in their respective webs

> of observables and assume the same (expectation) values. [Dieks et al. 2015, 209]

Now, given this situation, it might seem plausible to claim that the dual theories are really physically equivalent. Consider **RT**. On the internal view, we are led to say that 'area' really has the same meaning as 'entanglement entropy'. After all, in the theoretical structure that is supposed to matter on the internal view, viz. the empirical core, the two terms are related structurally in the same way to other terms elsewhere (sans a proportional constant). Given that the two theories is also stripped of all prior physical meaning, this structural identity suggests that the 'area' and 'entanglement entropy' are really describing the same quantities, despite their obvious non-isomorphism more generally (e.g. different equations in computing these quantities in their respective theories, the terms involved in calculating them, and so on). In other words, it seems that we are allowed to proclaim physical equivalence on this view.

If we do accept this third claim, we get physical and hence theoretical equivalence, and so the internal view does preclude the possibility of emergence: Theoretical equivalence effectively rules out any account of emergence. If the two dual theories are really just different formulations of one theory, then there is nothing for this new, unified, theory to emerge *from*: nothing can emerge from itself in any interesting way. Subsequently, a duality is supposed to *preclude* emergence on the internal view.

Agreed: physical equivalence entails theoretical equivalence, and theoretical equivalence rules out any sort of emergence. However, are we forced to adopt physical equivalence given the internal view? De Haro himself seems unclear on this point. Note the use of "can" in his characterization of the internal view above: "the two theories,

related by the duality, *can* describe the same physical quantities" [2017, 116, emphasis mine]. Are we supposed to believe that physical equivalence *can* hold, or that it *must* hold, on the internal view? In other words, since physical equivalence hangs on the third claim of leaving terms of the dual theories uninterpreted, *must* we adopt the third claim, or is it merely *possible*?

De Haro seems to suggest that theoretical, and hence physical, equivalence *must* hold, since he assumes the two dual theories to be 'two formulations of *one theory*' [emphasis mine]. However, later on, he suggests that physical equivalence merely *can* hold, when he considers an example of leaving dual theories uninterpreted beyond structural relations:

> For what might intuitively be interpreted as a 'length, a reinterpretation in terms of 'renormalisation group scale is now *available*.[5] [De Haro 2017, 116, emphasis mine]

The *availability* of an interpretative stance – in our case of **RT**, of interpreting bulk boundary surface area to be the same physical quantity as entanglement entropy – surely does not entail the *necessity* of the stance. Hence, there are two readings of the internal view: on the weak reading, we take the modal talk – e.g. a reinterpretation being 'available' or how we 'can' describe the same physical quantities – seriously, and on the strong reading we ignore the modal talk completely.

On the one hand, the claim that the internal view precludes emergence is not true on the weaker view. On this view, *if* we assume that the terms on both sides of the duality are uninterpreted, then there is no emergence; *but* this is not forced on us. In turn, this

[5]For context, though unmentioned in this paper, length and renormalisation group scale are also dual quantities in AdS/CFT.

makes the preclusion of emergence merely possible. However, this reading of the internal view does not rule out emergence as De Haro claims. I will thus assume that De Haro intends for us to take the strong reading of the internal view, which does claim that the terms of the both sides *are* uninterpreted.

However, we have not yet seen a compelling reason for accepting the claim that we *have to* see the terms of the dual theories as uninterpreted, and subsequently that physical equivalence *must* hold. *A fortiori* we are not obliged to accept the internal view.

Indeed, something is odd about the argument structure I mapped out: To establish the second claim of physical equivalence, we must establish the third claim, that we must discard anything beyond the empirical core and to leave the terms uninterpreted. However, to justify leaving the terms uninterpreted requires a convincing argument for assuming physical equivalence between the two theories to begin with! Otherwise, we have no reason to simply discard the 'excess' structure and leave the dual theories' terms uninterpreted.

Hence, further arguments are required to establish the third claim. Furthermore, if we discover that this argument is wanting, we shall then have reasons to reject the internal view.

4 De Haro's Argument

De Haro does provide an argument, which runs on the idea that two plausible commitments entails the internal view: the commitment that the dual theories are theories of the whole world in some suitably totalizing manner, and the commitment to "some form of structural realism" [2017, 116].

Let us begin by examining the two commitments. The first commitment implies that dual theories are theories of the whole world, in the sense that they are "both candidate descriptions of the same world" [Dieks et al. 2015, 14]. However, *prima facie* this is not true, since on one hand we have a theory of gravity/space-time geometry, while on the other we have a theory without (not to mention different dimensionalities). How can two theories, one describing something the other does not, both be about the same world? We can try to make this assumption intelligible by taking into account the translation manual between the two theories. Given the translation manual, we can claim that the **CFT** theory without gravity does describe gravity in a way. Consider **RT**: while the entanglement entropy described within **CFT** does not appear to describe space-time geometry *by itself*, the **CFT** plus the translation manual *and* **AdS** (in this case **RT**) *does* describe space-time geometry, albeit in a higher-dimensional space-time. When the entanglement described within the **CFT** changes, the boundary surface area in the **AdS**-side theory with gravity changes as well. Hence, by considering the translation manual given by the duality, the first commitment is made plausible.

The second commitment requires us to adopt some form of structural realism. Structural realism here can be understood loosely, since nothing turns on the particular account of structural realism we employ. Furthermore, De Haro himself does not specify precisely what he means by 'some form of' structural realism. As such, I will likewise adopt a loose notion of structural realism: I understand it to be the view that we should be (metaphysically or epistemically) committed only to the mathematical or formal structure of our theories, and this entails, among other things, that theoretical terms are to be defined in terms of their relations to other places or nodes in this formal structure.

Now, De Haro then claims that the two commitments entail the internal view:

> If [the two commitments] are met, it is impossible, in fact meaningless, to decide that one formulation of the theory is superior, since both theories are equally successful by all epistemic criteria one should apply. [De Haro 2017, 116]

Since he does not flesh out his argument in much detail, I attempt to reconstruct his argument in a plausible fashion: firstly, let us grant the two commitments. Do these commitments commit us to the conclusion that it is meaningless to differentiate between the two dual theories?

Dieks et al. [2015, 209] claims that given the first commitment, "it is no longer clear that there exists an 'external' point of view that independently fixes the meanings of terms in the two theories". However, I must admit I do not see why this is the case: as I explained above, the first commitment only makes sense *if* we understand both theories as having pre-determined meanings, and *then* relating them via the duality/translation manual. In other words, the first commitment is perfectly compatible with the external view.

For the remainder of this paper I focus on the second commitment instead. I think the second commitment *does* entail that differentiating the two theories is meaningless, *only if* we believe that one should be a structural realist (epistemically/metaphysically) only about the empirical core of the duality, discarding the 'excess structure' which made the two theories distinct structures to begin with. In other words, we want to say that this 'excess structure' was not physically significant to begin with: only the empirical core was relevant to physics. It seems that this is required to make sense of the claim that it is 'meaningless' to say that one formulation, e.g. the **CFT** side, is better than the

other, e.g. the **AdS** side. If structural realism commits us only to the empirical core of the dual theories, then accordingly there is really only one structure in question. Hence, it is meaningless to ask which structure is better (there is only one). If there is only one structure, then the internal view seems to hold: under a structural realist view, the terms of the dual theories are defined in terms of their places in the structure. Hence, within the empirical core's structure, the different terms of the dual theories really mean the same thing, and hence we get some version of the internal view.

Why should we, even as structural realists, commit ourselves only to the empirical core? The argument seems to me to be an epistemic one: we should believe that the structure relevant to the two theories given the duality must really be common to both theories because, as De Haro claims above, "both theories are equally successful" by all epistemic criteria we apply. If this is true then it seems we have no way of differentiating between the two theories, and the best explanation for this epistemic equivalence is to appeal to their being 'the same' in some way. The only thing in common between the dual theories is the empirical core, so we should take this to be what explains their epistemic equivalence. Everything else (i.e. the 'excess structure') can be discarded, since they are irrelevant differences. As such, structural realism should commit us only to the empirical core.

However, it is not clear that the dual theories are indeed epistemically equivalent. In a naive sense, they are epistemically equivalent if one takes 'epistemic' to be 'empirical' equivalence. Given the duality, i.e. formal and predictive equivalence, it is trivial that the two theories are also 'empirically' equivalent. However, I do not think such a notion of empirical equivalence *exhausts* the epistemic criteria for differentiating between scientific theories. Of course, one main desideratum for scientific theorizing is to provide

predictions, descriptions and explanations of phenomena. Beyond that, though, I contend that another desideratum of scientific theorizing is to look for ways to develop better scientific theories, be it a more unificatory theory, a more explanatory theory, and so on.

We see this in play when De Haro discusses the position/momentum duality in quantum mechanics: "this duality is usually seen as teaching us something new about the nature of reality: namely, that atoms are neither particles, nor waves. By analogy, it is to be expected that gauge/gravity dualities teach us something about the nature of spacetime and gravity" [2017, 117]. However, this is only possible *if* the two theories were not epistemically equivalent! If they were epistemically equivalent, then how could we learn anything new from one theory that we cannot already learn from another? If 'area' and 'entanglement entropy' really meant the same thing and had the same physical interpretation, how could we learn something new when we realize that area can be related (via **RT**) to quantum entanglement? Indeed, this criticism extends generally to the internal view: how can we learn anything new from a duality if the dual theories are just the 'same theory', and indeed are *uninterpreted* to begin with? We learn something new when two *different* things are related in a surprising way, *especially* when they are related to other quantities, on each side, in interesting ways; I do not see how we can learn something new when one and the same thing is related to itself.

Furthermore, the two theories are *not* epistemically equivalent when we consider the methodological concerns of physicists, who generally note that the **CFT** is well-understood, while the dual string theory of gravity is not. For example, Horowitz and Polchinski [2009] notes that we only approximately understand the gravitational theory, but the **CFT** has been developed to very precise degrees. Lin points out that:

> A dictionary is reasonably well developed in the direction of using classical gravity to study the **CFT**, but the converse problem how to organize the information in certain **CFT**'s into a theory of quantum gravity with a semi-classical limit is hardly understood at all. [2015, 11]

If both theories are equally successful by *all* epistemic criteria we have, then this situation should not appear. Rather, it seems that scientific practice is of the opinion that the two theories are, in fact, *not* epistemically equal: one is more successful than the other in terms of a variety of criteria, such as precision of calculation, ease of understanding, availability of a non-perturbative analysis, and so on. It is one reason why **AdS/CFT** is such an interesting area of research: it allows us to understand a hard-to-understand theory in terms of an easier-to-understand theory. Unless one is given arguments for why such criteria should *not* be epistemically relevant, the dual theories, I contend, are *not* epistemically equivalent.

Of course, one could assume that the *goal* or *ideal*, when we fully understand the translation manual, is to render both theories equally epistemically successful. However, this presumes that both sides *will* end up being just as easy to compute, or understand, and so on. Of course, if we do discover a more fundamental characterization of *why* the two dual theories are related by the duality as such, e.g. the sort of 'deeper' theory Rickles [2011, 2017] hopes for, then clearly we are entitled to the internal view since this 'deeper' theory will ideally explain why the dual theories, despite their apparent differences, can be seen as different facets of a single theory, just like how special relativity unified electromagnetism and made it plausible to understand both the electric and magnetic fields as facets of the 'deeper' Faraday tensor field. Right now, though,

there is no such theory in sight, making this point inadequate for supporting the internal view.

Given the foregoing, it is not clear there is epistemic equivalence: the epistemic argument does not hold. The upshot is that we are not compelled to provide an explanation for why the dual theories are epistemically equivalent to begin with (they are not), and hence we have no need to commit ourselves only to the common empirical core, *even* as structural realists, nor to think that differentiating the dual theories is meaningless.

Recall the oddity I pointed out in §3, though. The claim of physical equivalence hangs on leaving the dual theories uninterpreted, but this latter claim was itself motivated by physical equivalence. It was hoped, then, that the epistemic argument could provide independent motivation for adopting physical equivalence. Given my criticism of De Haro's additional argument, though, the circle returns, and leaves the two claims uncompelling. Hence, we should not adopt the internal view itself. Furthermore, my criticisms suggest that the dual theories are in fact *not* epistemically equivalent, and this suggests that the default stance is one where the two theories are not theoretically equivalent at all. Given the duality, the only way this can be so is to adopt the view that the dual theories are physically non-equivalent; in other words we should adopt the external view instead.

To conclude, given the dialectic set up by the interpretive fork, and the inadequacies of the internal view, I suggest that we adopt the external view instead.

5 The Way Forward

Let me end by commenting on the external view and the broader debate on whether there is emergence given a duality (§1). In §3 we have seen how the internal view precludes emergence simply because there are no two distinct theories to speak of: we merely have two ways of looking at a single theory. This in turn swiftly rules out any talk of emergence. The external view, though, does not rule out emergence quite so easily, and there is some leeway to speak of emergence since we *do* have two distinct theories which are, as Teh noted, generically *not* isomorphic to one another. However, given the formal and predictive equivalences demanded by a duality relation, a duality relation is symmetric, and so there is nothing within a duality that will formally broker the asymmetry between two theories we often associate with emergence. One way to do so, as Teh (2013) suggests, is to introduce a claim of relative fundamentality, i.e. which theory is 'more fundamental' than another, is required to break the symmetry and provide us with the required asymmetry for emergence. While the external view does not entail this, it does not rule it out either. Hence, the external view does not preclude emergence; instead, it directs attention about emergence and duality away from the interpretative fork, onto whether and how one can make claims about relative fundamentality in the context of dualities. Alas, this requires much more attention than I can afford here: I leave it for another day.

References

Dieks, D., J. van Dongen, and S. D. Haro (2015). Emergence in Holographic Scenarios for Gravity. *Studies in the History and Philosophy of Modern Physics 52*, 203–216. 10.1016/j.shpsb.2015.07.007.

Fraser, D. (2017). Formal and Physical Equivalence in Two Cases in Contemporary Quantum Physics. *Studies in the History and Philosophy of Modern Physics 59*, 30–43. 10.1016/j.shpsb.2015.07.005.

Haro, S. D. (2017). Dualities and Emergent Gravity: Gauge/Gravity Duality. *Studies in the History and Philosophy of Modern Physics 59*, 109–125. DOI: 10.1016/j.shpsb.2015.08.004.

Haro, S. D., N. Teh, and J. Butterfield (2017). Comparing Dualities and Gauge Symmetries. *Studies in the History and Philosophy of Modern Physics 59*, 68–80. DOI: 10.1016/j.shpsb.2016.03.001.

Horowitz, G. and J. Polchinski (2009). Gauge/gravity duality. In D. Oriti (Ed.), *Approaches to quantum gravity: Toward a new understanding of space time and matter*, pp. 169–186. Cambridge: Cambridge University Press. arXiv:gr-qc/0602037.

Raamsdonk, M. V. (2010). Building up spacetime with quantum entanglement. *General Relativity and Gravitation 42*(10), 2323–2329. 10.1007/s10714-010-1034-0.

Rickles, D. (2011). A Philosopher Looks at Dualities. *Studies in the History and Philosophy of Modern Physics 42*(1), 54–67. DOI: 10.1016/j.shpsb.2010.12.005.

Rickles, D. (2017). Dual Theories: 'Same but Different or 'Different but Same? *Studies in the History and Philosophy of Modern Physics 59*, 62–67. 10.1016/j.shpsb.2015.09.005.

Ryu, S. and T. Takayanagi (2006). Holographic Derivation of Entanglement Entropy from AdS/CFT. *Phys. Rev. Lett 96*(18). 10.1103/PhysRevLett.96.181602.

Seiberg, N. (2007). Emergent spacetime. In D. Gross, M. Henneaux, and A. Sevrin (Eds.), *The Quantum Structure of Space and Time*, pp. 163–178. Singapore: World Scientific. DOI: 10.1142/9789812706768_0005.

Teh, N. (2013). Holography and Emergence. *Studies in the History and Philosophy of Modern Physics 44*(3), 300–311. DOI: 10.1016/j.shpsb.2013.02.006.

Causal Discovery and the Problem of Psychological Interventions

PSA 2018, Seattle

Markus Eronen

University of Groningen

m.i.eronen@rug.nl

Abstract

Finding causes is a central goal in psychological research. In this paper, I argue that the search for psychological causes faces great obstacles, drawing from the interventionist theory of causation. First, psychological interventions are likely to be both fat-handed and soft, and there are currently no conceptual tools for making causal inferences based on such interventions. Second, holding possible confounders fixed seems to be realistically possible only at the group level, but group-level findings do not allow inferences to individual-level causal relationships. I also consider the implications of these problems, as well as possible ways forward for psychological research.

1. Introduction

A key objective in psychological research is to distinguish causal relationships from mere correlations (Kendler and Campbell 2009; Pearl 2009; Shadish and Sullivan 2012). For example, psychologists want to know whether having negative thoughts is a cause of anxiety instead of just being correlated with it: If the relationship is causal, then the two are not just spuriously hanging together, and intervening on negative thinking is actually one way of reducing anxiety in patients suffering from anxiety disorders. However, to what extent is it actually possible to find psychological causes? In this paper, I will seek an answer this question from the perspective of state-of-the-art philosophy of science.

In philosophy of science, the standard approach to causal discovery is currently interventionism, which is a very general and powerful framework that provides an account of the features of causal relationships, what distinguishes them from mere correlations, and what kind of knowledge is needed to infer them (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009, Woodward 2003, 2015b; Woodward & Hitchcock 2003). Interventionism has its roots in Directed Acyclic Graphs (DAGs), also known as causal Bayes nets, which are graphical representations of causal relationships based on conditional independence relations (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009). More recently, James Woodward has developed interventionism into a full-blown philosophical account of causation, which has become popular in philosophy and the sciences. Several authors have also argued that interventionism adequately captures the role of causal thinking and reasoning in psychological research (Campbell 2007; Kendler and Campbell 2009; Rescorla forthcoming; Woodward 2008).

Based on interventionism, I will argue in this paper that the discovery of psychological causes faces great obstacles. This is due to problems in performing psychological interventions and deriving interventionist causal knowledge from psychological data.[1] Importantly, my focus is not on the existence or possibility of psychological causation, but on the *discovery* of psychological causes, which is a topic that has so far received little attention in philosophy.[2] Although I rely on interventionism, my arguments are based on rather general principles of causal inference and reasoning in science, and will thus apply to any other theory of causation that does justice to such principles.

The focus in this paper will be on the discovery *individual-level* (or within-subject) causes, not *population-level* (or between-subjects) causes. The first refers to causal relationships that hold for a particular individual: for example, John's negative thoughts cause John's problems of concentration. The latter refers to causal relationships that obtain in the population as a whole: for example, negative thoughts cause problems of concentration in a population of university students. It widely thought that ultimate goal of causal inference is to find individual-level causes, and that a population-level causal relationship should be seen as just an average of individual-level causal relationships (Holland 1986): For example, the causal relationship between negative thoughts and problems of concentration in a population of university students is only interesting insofar as it *also* applies to at least some of the individual students in the

[1] See Eberhardt (2013; 2014) for different (and domain-independent) problems for interventionist causal discovery.
[2] There is an extensive debate on the question whether interventionism vindicates non-reductive psychological causation by providing a solution the causal exclusion problem (e.g., Baumgartner 2009, Eronen 2012, Raatikainen 2010, Woodward 2015). I will sidestep this debate here, as my focus is not on the existence of non-reductive psychological causation, but on the discovery of psychological causes, be they reducible or not.

population.[3] Thus, in this paper I will discuss population-level causal relationships only when they are relevant to discovering individual-level causes.

Importantly, the distinction between population-level and individual-level causation is different from the distinction between type and token causation, even though the two distinctions are sometimes mixed up in the philosophical literature (see also Illari & Russo 2014, ch. 5). Token causation refers to causation between two actual events, whereas type causation refers to causal relationships that hold more generally. Individual-level causes can be either type causes or token causes. An example of an individual and type causal relationship would be that John's pessimistic thoughts cause John's problems of concentration: This is a general relationship between two variables, and not a relationship between two actual events. An example of an individual and token causal relationship would be that John's pessimistic thoughts before the exam on Friday at 2 pm caused his problems of concentration in the exam. As interventionism is a type-level theory of causation, and the aim of psychological research is primarily to discover regularities, not explanations to particular events, in this paper I will only discuss the discovery of type (individual) causes.

The structure of this paper is as follows. I will start by giving a brief introduction to interventionism, and then turn to problems of interventionist causal inference in psychology: First, to problems related to psychological interventions (section 2), and then to problems arising from the requirement to "hold fixed" possible confounders (section 3). After this, I will consider the possibility of the inferring psychological causes without interventions (section 4). In the last

[3] It has been argued that population-level (between-persons) causal relationships can also be real without applying to any individual (Borsboom, Mellenbergh, and van Heerden 2003). However, also those who believe in these kind of population-level causes agree that discovering individual causes is an important goal as well.

section, I discuss ways forward and various implications that my arguments have for psychology and its philosophy.

2. Interventionism

Interventionism is a theory of causation that aims at elucidating the role of causal thinking in science, and defining a notion of causation that captures the difference between causal relationships and mere correlations (Woodward 2003). Thus, the goal of interventionism is to provide a methodologically fruitful account of causation, and *not* to reduce causation to non-causal notions or analyse the metaphysical nature of causation (Woodward 2015b). In a nutshell, interventionist causation is defined as follows:

X is a cause of Y (in variable set **V**) if and only if it is possible to *intervene* on X to change Y when all other variables (in **V**) that are not on the path from X to Y are *held fixed* to some value (Woodward 2003).

Thus, in order to establish that X is a cause of Y, we need evidence that there is some way of intervening on X that results in a change in Y, when off-path variables are held fixed.[4] Importantly, it is not necessary to actually perform an intervention: What is necessary is knowledge on what *would* happen if we *were* to make the right kind of intervention.

[4] More precisely, this is the definition for a *contributing* cause. X is a *direct* cause of Y if and only if it is possible to intervene on X to change Y when all other variables (in **V**) are held fixed to some value (Woodward 2003). Thus, the definition of a contributing cause allows there to be other variables on the causal path between X and Y, whereas the definition of a direct cause does not. This does not reflect any substantive metaphysical distinction, as the question whether X is a direct or contributing cause is relative to what variables are included in the variable set. Importantly, notion of a contributing cause is *not* relative to a variable set in any strong sense – if X is a cause of Y in some variable set, then X will be a cause of Y in all variable sets where X and Y appear (Woodward 2008b). This is because the definition of an intervention is not relativized to a variable set.

The notion of an intervention plays a fundamental role in the account, and is very specifically defined. Here is a concise description of the four conditions that an intervention has to satisfy (Woodward 2003).

Variable *I* is an intervention variable for X with respect to Y if and only if:

(I1) *I causes* the change in X;

(I2) The change in X is *entirely* due to *I* and not any other factors;

(I3) *I* is not a cause of Y, or any cause of Y that is not on the path from X to Y;

(I4) *I* is *uncorrelated* with any causes of Y that are not on the path from X to Y.

The rationale behind these conditions is that if the intervention does not satisfy them, then one is not warranted to conclude that the change in Y was (only) due to the intervention on X. Thus, in simpler terms, the intervention should be such that it changes the value of the target variable *X* in such a way that the change in *Y* is *only* due to the change in *X* and not any other influences (Woodward 2015b). For example, if the intervention is correlated with some other cause of Y, say Z, that is not on the path from X to Y (violating I4), then the change in Y may have been (partly) due to Z, and not just due to X. Following standard terminology in the literature, I will call interventions that satisfy the criteria I1-I4 *ideal* interventions. I will now go through various problems in performing ideal interventions in psychology, starting from problems related to conditions I2 and I3 (section 3), and then turn to problems related to I4 and the "holding fixed" part of the definition of causation (section 4).

3. Psychological interventions

Before discussing psychological interventions, an important distinction needs to be made: The distinction between relationships where (1) the cause is *non-psychological*, and the *effect* is psychological, and (2) where the *cause* (and possibly also the effect) is *psychological*.[5] A large proportion (perhaps the majority) of experiments in psychology involve relationships of the first kind: The intervention targets a non-psychological variable (X) such as medication vs. placebo, therapy regime vs. no therapy, or distressing vs. neutral video, and the psychological effect of the manipulation of this non-psychological variable is tracked. In other words, the putative causal relation is between a non-psychological cause variable (X) and a psychological effect variable (Y). In these cases, it is possible to do (nearly) ideal interventions on the putative cause variable (X) by ensuring that the change in X was caused (only) by the intervention, that the intervention did not change Y directly, and that it was uncorrelated with other causes of Y. It is of course far from trivial to make sure that these conditions were satisfied, but as the variables intervened upon are non-psychological, making the right kinds of interventions is in principle not more difficult than in other fields. As regards the psychological effect variable (Y), there is no need to intervene on it; it is enough to measure the change in Y (which, again, is far from trivial, but faces just the usual problems in psychological measurement, which will be discussed below). The fact that many psychological experiments involve this kind of causal relationships may have contributed to the recent optimism on the prospects of interventionist causal inference in psychology.

[5] The line between psychological and non-psychological variables is likely to be blurry. However, for the present purposes it is not crucial where exactly the line should be drawn: My arguments apply to cases where it is clear that the cause variable is psychological (such as the examples in the main text), and such cases abound in psychological research.

However, psychological research also often concerns relationships of the second kind, that is, relationships where the *cause* is psychological. This is, for example, the case when the aim is to uncover psychological mechanisms that explain cognition and behavior (e.g., Bechtel 2008, Piccinini & Craver 2011), or to find networks of causally interacting emotions or symptoms (e.g., Borsboom & Cramer 2013). The reason why these relationships are crucially different from relationships of the first kind is that now the variable intervened upon is psychological, so the conditions on interventions now have to be applied to psychological variables.

Ideal interventions on psychological variables are rarely if ever possible. One reason for this has been extensively discussed by John Campbell (2007): Psychological interventions seem to be "soft", meaning that the value of the target variable *X* is not completely determined by the intervention (Eberhardt & Scheines 2007; see also Kendler and Campbell 2009; Korb and Nyberg 2006). In other words, the intervention does not "cut off" all causal arrows ending at *X*. As a non-psychological example, when studying shopping behaviour during one month by intervening on income, an ideal intervention would fully determine the exact income that subjects have that month, whereas simply giving the subjects an *extra* 5000€ would count as a soft intervention (Eberhardt & Scheines 2007). Similarly, if we intervene on John's psychological variable *alertness* by shouting "WATCH OUT!", this does not completely cut off the causal contribution of other psychological variables that may influence John's *alertness*, but merely adds something on top of those causal contributions (Campbell 2007). As most or all interventions on psychological variables are likely to be soft, Campbell proposes that we should simply allow such soft interventions in the context of psychology. Campbell argues that these kind of interventions can still be informative and indicative of causal relationships (Campbell

2007), and this conclusion is supported by independent work on soft interventions in the causal modelling literature (e.g., Eberhardt & Scheines 2007; Korb and Nyberg 2006).

However, the problem of psychological interventions is not solved by allowing for soft interventions. There is a further, equally important reason why interventions on psychological variables are problematic: Psychological interventions typically *change several variables simultaneously*. For example, suppose we wanted to find out whether *pessimistic thoughts* cause *problems in concentration*. In order to do this, we would have to find out what would happen to *problems in concentration* if we were to intervene just on *pessimistic thoughts* without perturbing other psychological states with the intervention. However, how could we intervene on *pessimistic thoughts* without changing, for example, *depressive mood* or *feelings of guilt*? As an actual scientific example, consider a network of psychological variables that includes, among others, the items *alert*, *happy*, and *excited* (Pe et al. 2015). How could we intervene on just one of those variables without changing the others?

One reason why performing "surgical" interventions that only change one psychological variable is so difficult is that there is no straightforward way of manipulating or changing the values of psychological variables (as in, for example, electrical circuits). Interventions in psychology have to be done, for example, through verbal information (as in the example of John above) or through visual/auditory stimuli, and such manipulations are not precise enough to manipulate just one psychological variable. Also state-of-the-art neuroscientific methods such as Transcranial Magnetic Stimulation affect relatively large areas of the brain, and are not suited for intervening on specific psychological variables. Currently, and in the foreseeable future, there is no realistic

way of intervening on a psychological variable without at the same time perturbing some other psychological variables.

Thus, it is likely that most or even all psychological interventions do not just change the target variable X, but also some other variable(s) in the system. In the causal modelling literature, interventions of this kind have been dubbed *fat-handed*[6] interventions (Baumgartner and Gebharter 2016; Eberhardt & Scheines 2007; Scheines 2005). For example, an intervention on pessimistic thoughts that also immediately changes depressive mood is fat-handed. Fat-handed interventions have been recently discussed in philosophy of science, but mainly in the context of mental causation and supervenience (e.g., Baumgartner and Gebharter 2016, Romero 2015), and the fact that psychological interventions are likely to be systematically fat-handed (for reasons unrelated to supervenience) has not yet received attention.

An additional complication is that it is difficult check what a psychological intervention precisely changed, and to what extent it was fat-handed (and soft). In fields such as biology or physics there are usually several independent ways of measuring a variable: for example, temperature can be measured with mercury thermometers or radiation thermometers, and the firing rate of a neuron can be measured with microelectrodes or patch clamps. However, measurements of psychological variables, such as emotions or thoughts, are based on self-reports, and there is no further independent way of verifying that these reports are correct. Moreover, only a limited number of psychological variables can be measured at a given time point, so an intervention may always have unforeseen effects on unmeasured variables.

[6] According to Scheines (2005), this term was coined by Kevin Kelly.

Why are fat-handed interventions so problematic for interventionist causal inference? The reason becomes clear when looking at condition I3: The intervention should not change any variable Z that is on a causal pathway that leads to Y (except, of course, those variables that are on the path between X and Y). If the causal structure of the system under study is known, as well as the changes that the intervention causes, then this condition can sometimes be satisfied even the intervention was fat-handed. However, in the context of intervening on psychological variables, neither the causal structure nor the exact effects of the interventions are known. Thus, when the intervention is fat-handed, it is not known whether I3 is satisfied or not, and in many cases it is likely to be violated. In other words, we cannot assume that the intervention was an unconfounded manipulation of X with respect to Y, and cannot conclude that X is a cause of Y.

4. The Problem of "Holding Fixed"

The next problem that I will discuss is related to the last part of the definition of interventionist causation: X is a cause of Y (in variable set **V**) if and only if it is possible to intervene on X to change Y *when all other variables (in **V**) that are not on the path from X to Y are held fixed to some valu*e. The motivation for this requirement is to make sure that the change in Y is really due to the change X, and not due to some other cause of Y. To a large extent, this is just another way of stating what is already expressed in the definition of an intervention, in conditions I3 and I4: The intervention should not be confounded by any cause of Y that is not on the path between X and Y.[7] In the previous section, we saw that fat-handed interventions pose a challenge for

[7] In recent publications, Woodward often gives a shorter definition of causation that does not include the "holding fixed" part, for example: "X causes Y if and only if under some interventions on X (and possibly other variables) the value of Y changes" (Woodward 2015). This is understandable, as the definition of intervention already contains conditions I3 and I4, which effectively imply holding fixed potential causes of Y that are correlated with the intervention and are not on the path from X to Y. However, there are also good reasons why the full definition has to

satisfying this condition. However, as I will now show, it is problematic in psychology also for more general reasons.

In psychology, it is impossible to hold psychological variables fixed in any concrete way: We cannot "freeze" mental states, or ask an individual to hold her thoughts constant. Thus, the same effect has to be achieved indirectly, and the gold standard for this is Randomized Controlled Trials (RCTs) (Woodward 2003, 2008). RCTs have their origin in medicine, but are widely used in psychology and the social sciences as well (Clarke et al. 2014; Shadish, Cook and Campbell 2002; Shadish and Sullivan 2012). The basic idea of RCTs is to conduct a trial with two groups, the test group and the control group, which are as similar to one another as possible, but the test group receives the experimental manipulation and the control group does not. If the groups are large enough and the randomization is done correctly, any differences between the groups should be only due to the experimental manipulation. If everything goes well, this in effect amounts to "holding fixed" all off-path variables.

However, this methodology has an important limitation that has been overlooked in the literature on interventionism. As the effect of "holding fixed" is based on the difference between the groups as wholes, it only applies at the level of the group, and not at the level of individuals. For this reason, results of RCTs hold for the study population as a whole, but not necessarily for particular individuals in the population (cf. Borsboom 2005, Molenaar & Campbell 2009). For example, if we discover that pessimistic thoughts are causally related to problems of

include the second component as well. For example, consider a situation where we intervene on X with respect to Y, and Y changes, but this change is fully due to a change in variable Z, which is a cause of Y that is *uncorrelated* with the intervention variable. In this situation, without the "holding fixed" requirement we would falsely conclude that X is a cause of Y.

concentration in the population under study, it does not follow that this causal relationship holds in John, Mary, or any other specific individual in the population. This is related to the "fundamental problem of causal inference" (Holland 1986): Each individual in the experiment can belong to only one of the two groups (control or test group), and therefore cannot act as a "control" for herself, so only an average causal effect can be estimated. What this implies for causal inference in psychology is that when a causal relationship is discovered through an RCT, we cannot infer that this relationships holds for any specific individual in the population (see also Illari & Russo 2014, ch. 5).

This does not mean that the population-level findings based on RCTs are uninformative or useless. The point is rather that we currently have no understanding of when, to what extent and under what circumstances they also apply to the individuals in the population. This of course applies also to other fields where RCTs are used, such the biomedical sciences. Indeed, especially in the context of personalized medicine, the fact that RCTs are as such not enough to establish individual-level causal relationships has recently become a matter of discussion (e.g., de Leon 2012).

It might be tempting to simply look at the data more closely and find those individuals for whom the intervention on X actually corresponded with a change in Y. However, it would be a mistake to conclude that in those individuals the change in Y was caused by X. It might very well have been caused by some other cause of Y, as possible confounders were only held fixed at the group level, not at the individual level.[8] Thus, in RCTs possible confounders can only be held fixed at

[8] Would it be possible for a causal relationship to hold at the population level, but not for any individual in the population? Probably not, if the relationship is genuine: Weinberger (2015) has argued that there has to be at least *one* individual in the population for whom the relationships holds. However, in the context of discovery, it is

the group level, and this does not warrant causal inferences that apply to specific individuals. This is further limitation to interventionist causal inference in psychology.

5. Finding psychological causes without interventions

One possible response to the concerns raised in the previous two sections is that interventionism does not require that interventions are actually performed: As briefly mentioned in section 2, what is necessary is to know what *would* happen if we *were* to perform the right kinds of interventions. In other words, in order to establish that X is a cause of Y, it is enough to know that if we *were* to intervene on X with respect to Y (while holding off-path variables fixed), then Y *would* change. For example, it is beyond doubt that the gravitation of the moon causes the tides, even though no one has ever intervened on the gravitation of the moon to see what happens to the tides, and such an intervention would be practically impossible (Woodward 2003). Similarly, it could be argued that even though it is practically impossible to do (ideal) interventions on psychological variables, the knowledge on the effects of interventions could be derived in some other way. Let us thus consider to what extent this could be possible.

The state-of-the-art method for deriving (interventionist) causal knowledge when data on interventions is not available is *Directed Acyclic Graphs (DAGs)*, which were briefly mentioned in the introduction (see also Malinsky & Danks 2018, Pearl 2000, Spirtes, Glymour and Scheines 2000, Spirtes & Zhang 2016). Causal discovery algorithms based on DAGs take purely

possible that a causal *finding* at the population level is just an artefact of heterogeneous causal structures at the individual level, and therefore does not apply to any individual in the population.

observational data as input, and based on conditional independence relations, find the causal graph that best fits the data. In principle, these algorithms can be used for psychological data, with the aim of discovering causal relationships between psychological variables.

However, even though these algorithms do not require experimental data, they do require data from which conditional independence relations can be reliably drawn, and they (implicitly) assume that the variables that are modelled are independently and surgically manipulable (Malinsky & Danks 2018). In contrast, as should be clear from the above discussion, measurements of psychological variables typically come with a great deal of uncertainly, and it is not clear to what extent they can be independently manipulated. Moreover, causal discovery algorithms standardly assume *causal sufficiency*, that is, that there are no unmeasured common causes that could affect the causal structure (Malinsky & Danks 2018; Spirtes & Zhang 2016). The reason for this is that if two or more variables in the variable set have unmeasured common causes, then the inferences concerning the causal relationships between those variables will be either incorrect or inconclusive. However, missing common causes is likely the norm rather than the exception when it comes to psychological variables. For example, if the variable set consists of, say, 16 emotion variables, how likely is it that *all* relevant emotion variables have been included? And even if all emotion variables that are common causes to other emotion variables are included, is it plausible to assume that there are no further cognitive or biological variables that could be common causes to some of the emotion variables? As similar questions can be asked for any context involving psychological variables, causal sufficiency is a very unrealistic assumption for psychological variable sets.

For these reasons, psychological data sets are rather ill-suited for causal discovery algorithms, and these algorithms cannot be treated as reliable guides to interventionist causal knowledge in psychology. It is likely that the problems of psychological interventions discussed in the previous sections are not just practical problems in carrying out interventions, but reflect the immense complexity of the system under study (the human mind-brain), and therefore cannot be circumvented by just using non-experimental data (see, however, section 7 for a different approach).

6. Psychological interventions: A summary

To summarize, what I have argued so far is that interventionist causal inference in psychology faces several obstacles: (1) Psychological interventions are typically *both* fat-handed *and* soft: They change several variables simultaneously, and do not completely determine the value(s) of the variable(s) intervened upon. It is not known to what extent such interventions give leverage for causal inference. (2) Due to the nature psychological measurement, the degree to which a psychological intervention was soft and fat-handed, or more generally, what the intervention in fact did, is difficult to reliably estimate. (3) Holding fixed possible confounders is only possible at the population level, not at the individual level, and it is not known under what conditions population-level causal relationships also apply to individuals. (4) Causal inference based on data without interventions requires assumptions that are unrealistic for psychological variable sets. Taken together, these issues amount to a formidable challenge for finding psychological causes.[9]

[9] Baumgartner (2009, 2012, 2018) has argued that mental-to-physical supervenience makes it impossible to satisfy the Woodwardian conditions on interventions, and that if interventionism is modified to accommodate supervenience relationships (as in Woodward 2015), the result is that any causal structure with a psychological

7. Discussion

Although various metaphysical and conceptual issues related to psychological causation have been extensively discussed in philosophy of science, little attention has been paid to the *discovery* of psychological causes. In this paper, I have contributed to filling this lacuna, by discussing the search for psychological causes in the framework of the interventionist theory of causation. The upshot is that finding individual psychological causes faces daunting challenges. The problems in holding fixed confounders and performing interventions need to be taken into account when trying to establish a psychological causal relationship, or when making claims about psychological causes.

However, I do not want to argue that finding psychological causes is *impossible*, or that researchers should stop looking for psychological causes. Rather, my aim is to contribute to getting a better understanding of the limits of finding causes in psychology, and the challenges involved. This can also lead to positive insights regarding causal inference in psychology. One such insight is that more attention should be paid to *robust inference* or *triangulation*. Often when individual methods or sources of evidence are insufficient or unreliable, as is the case here, what is needed is a more holistic approach. A widespread (though not uncontroversial) idea in philosophy of science is that evidence from several independent sources can lead to a degree of confidence even if the sources are individually fallible and insufficient (Eronen 2015, Kuorikoski

cause becomes empirically indistinguishable from a corresponding structure where the psychological variable is epiphenomenal. If this reasoning is correct, it leads to a further (albeit more theoretical) problem for interventionist causal inference: Any empirical evidence for a causal relationships with a psychological cause is equally strong evidence for a corresponding epiphenomenal structure, and it is not clear which structure should be preferred and on what grounds.

& Marchionni 2017, Munafo & Smith 2017, Wimsatt 1981, 1994/2007). For example, there is no single method or source of evidence that would be individually sufficient to establish that the anthropogenic increase in carbon dioxide is the cause for the rise in global temperature, but there is so much converging evidence from many independent sources that scientists are confident that this causal relationship exists. Similarly, evidence for a psychological causal relationship could be gathered from many independent sources: Several different (soft and fat-handed) interventions involving different variables, multilevel models based on time-series data, single-case observational studies, and so on.[10] If they all point towards the same causal relationships, this may lead to a degree of confidence in the reality of that relationship. However, how this integration of evidence would exactly work, and whether it can actually lead to sufficient evidence for psychological causal relationships, are open questions.

A related point is that psychological research can also make substantive progress *without* establishing causal relationships. Often important discoveries in psychology have not been discoveries of causal relationships, but rather discoveries of robust *patterns* or *phenomena* (Haig 2012, Rozin 2001, Tabb and Schaffner 2017). Consider, for example, the celebrated discovery that people often do not reason logically when making statistical predictions, but rely on shortcuts, for example, grossly overestimating the likelihood of dying in an earthquake or terror attack (Kahneman & Tversky 1973). In other words, when we reason statistically, we often rely on heuristics that lead to biases. The discovery of this phenomenon had nothing to with methods of causal inference (Kahneman and Tversky 1973), and its significance is not captured by describing causal relationships between variables. In fact, the causal mechanisms underlying the

[10] See also Peters, Bühlmann, & Meinshausen (2016), who present a formal model for inferring causal relationships based on their stability under different kinds of (non-ideal) interventions.

heuristics and biases of reasoning are still unknown. Similar examples abound in psychology: Consider, for example, groupthink or inattentional blindness. Of course, there are likely to be causal mechanisms that give rise to these phenomena, but the phenomena are highly relevant for theory and practice even when we know little or nothing about those underlying mechanisms (which is the current situation). This, in combination with the challenges discussed in this paper, suggests that (philosophy of) psychology might benefit from reconsidering the idea that discovering causal relationships is central for making progress in psychology.

Finally, one might wonder whether the problems I have discussed here are restricted to just psychology. Indeed, I believe that the arguments I have presented are more general, and apply to any other fields where there are similar problems with soft and fat-handed interventions and controlling for confounders. There is probably a continuum, where psychology is close to one end of the continuum, and at the other end we have fields where ideal interventions can be straightforwardly performed and variables can be easily held fixed, such as engineering science. Fields such as economics and political science are probably close to where psychology is, as they also face deep problems in making (ideal) interventions and measuring their effects. Same holds for neuroscience, at least cognitive neuroscience: The problems of soft and fat-handed interventions and holding variables fixed apply just as well to brain areas as to psychological variables (see also Northcott forthcoming). Thus, appreciating the challenges I have discussed here and considering possible reactions to them could also benefit many other fields besides psychology.

To conclude, I have argued in this paper that there are several serious obstacles to the discovery of psychological causes. As it is widely assumed in both psychology and its philosophy that the discovery of causes is a central goal, these obstacles need to be explicitly discussed, taken into account, and studied further.

References

Baumgartner, M. (2013). Rendering Interventionism and Non-Reductive Physicalism Compatible. *dialectica* 67: 1-27.

Baumgartner, M. (2018). The Inherent Empirical Underdetermination of Mental Causation. *Australasian Journal of Philosophy.*

Baumgartner, M and Gebharter, A. (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science* 67: 731-756.

Borsboom, Dennny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press

Borsboom, Denny and Anqelique O. Cramer. 2013. "Network analysis: an integrative approach to the structure of psychopathology." *Annual review of clinical psychology* 9: 91-121.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203.

Campbell, John. 2007. "An interventionist approach to causation in psychology." In: A. Gopnik & L. Schulz (eds.) *Causal Learning. Psychology, Philosophy, and Computation*. Oxford: Oxford University Press, 58–66.

Chirimuuta, Mazviita. Forthcoming. “Explanation in Computational Neuroscience: Causal and Non-causal.” *British Journal for the Philosophy of Science*. DOI:https://doi.org/10.1093/bjps/axw034

Clarke, B., Gillies, D., Illari, P, Russo, F., & Williamson, J. 2014. “Mechanisms and the evidence hierarchy.” *Topoi* 33: 339-360.

de Leon, J. (2012). Evidence-based medicine versus personalized medicine: are they enemies? *Journal of clinical psychopharmacology*, *32*(2), 153-164.

Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science*, *80*(5), 684-696.

Eberhardt, F. (2014). Direct causes and the trouble with soft interventions. *Erkenntnis*, *79*(4), 755-777.

Eberhardt, Frederick and Richard Scheines. 2007. “Interventions and causal inference.” *Philosophy of Science* 74: 981–995.

Eronen, Markus. Forthcoming. “Interventionism for the Intentional Stance: True Believers and Their Brains.” *Topoi*.

Hamaker, Ellen L. 2011. “Why researchers should think **“**within-person.**””** In M. R. Mehl, & T. A. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Kahneman, Daniel and Amos Tversky. 1973. “On the psychology of prediction.” *Psychological Review* 80: 237-251.

Kendler, Kenneth S. and John Campbell. 2009. Interventionist causal models in psychiatry: repositioning the mind-body problem. *Psychological Medicine* 39: 881-887.

Korb, K. B., & Nyberg, E. 2006. "The power of intervention." *Minds and Machines* 16*:* 289-302.

Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science, 83*, 227-247.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470.

Menzies, Peter. 2008. "The exclusion problem, the determination relation, and contrastive causation." In J. Hohwy & J. Kallestrup (Eds.) *Being Reduced* (pp. 196-217). Oxford: Oxford University Press.

Molenaar, Peter and Cynthia Campbell. 2009. "The new person-specific paradigm in psychology." *Current Directions in Psychological Science* 18: 112-117.

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature* 553, 399-401

Northcott, R. (forthcoming). Free will is not a testable hypothesis. *Erkenntnis*.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... & Kuppens, P. 2015. "Emotion-network density in major depressive disorder." *Clinical Psychological Science*, *3*(2), 292-300.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, Judea. 2009. "Causal inference in statistics: An overview." *Statistics surveys* 3: 96-146.

Pearl, Judea. 2014. "Comment: understanding simpson's paradox." *The American Statistician* 68: 8-13.

Peters, J. , Bühlmann, P. and Meinshausen, N. (2016), Causal inference by using invariant prediction: identification and confidence intervals. J. R. Stat. Soc. B, 78: 947-1012. doi:10.1111/rssb.12167

Rescorla, Michael. Forthcoming. "An interventionist approach to psychological explanation." *Synthese*.

Reutlinger, Alexander and Juha Saatsi (eds.). 2017. *Explanation Beyond Causation*. Oxford: Oxford University Press.

Romero, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese*, *192*(11), 3731-3755.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2-14.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science*, *72*(5), 927-940.

Shadish W. R., Cook T. D. and Campbell D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin; Boston.

Shadish, W. R., & Sullivan, K. J. 2012. "Theories of causation in psychological science." In H. Cooper et al. (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 23-52). Washington, DC: American Psychological Association.

Shapiro, Lawrence. 2010. "Lessons from causal exclusion." *Philosophy and Phenomenological Research*, 81, 594-604.

Shapiro, Lawrence. 2012. "Mental manipulations and the problem of causal exclusion." *Australasian Journal of Philosophy*, 90, 507–524.

Shapiro, Lawrence and Elliott Sober. 2007. “Epiphenomenalism: the dos and the don’ts.” In G. Wolters & P. Machamer (Eds.) *Thinking about causes: from Greek philosophy to modern physics* (pp. 235–264). Pittsburgh, PA: University of Pittsburgh Press.

Spirtes, Peter, Glymour, Clark and Richard Scheines. 2000. *Causation, prediction, and search*. New York: Springer.

Tabb, K., & Schaffner, K. F. (2017). Causal pathways, random walks and tortuous paths: Moving from the descriptive to the etiological in psychiatry. In: Kendler, K. S., & Parnas, J. (Eds.) *Philosophical Issues in Psychiatry IV: Nosology* (pp. 342-360). Oxford: Oxford University Press.

Weinberger, Naftali. 2015. “If intelligence is a cause, it is a within-subjects cause.” *Theory & Psychology*, 25(3), 346-361.

Woodward, James. 2003. *Making things happen. A theory of causal explanation.* Oxford: Oxford University Press.

Woodward, James. 2008. “Mental causation and neural mechanisms.” In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: new essays on reduction, explanation, and causation*. Oxford: Oxford University Press: 218-262

Woodward, James. 2015a. “Interventionism and causal exclusion.” *Philosophy and Phenomenological Research* 91, 303-347.

Woodward, James. 2015b. “Methodology, ontology, and interventionism.” *Synthese* 192, 3577-3599.

Woodward, James & Christopher Hitchcock. 2003. “Explanatory Generalizations, Part I: A Counterfactual Account.” *Noûs* 37(1): 1–24.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

Why Replication is Overrated

Current debates about the replication crisis in psychology take it for granted that direct replication is valuable and focus their attention on questionable research practices in regard to statistical analyses. This paper takes a broader look at the notion of replication as such. It is argued that all experimentation/replication involves individuation judgments and that research in experimental psychology frequently turns on probing the adequacy of such judgments. In this vein, I highlight the ubiquity of conceptual and material questions in research, and I argue that replication is not as central to psychological research as it is sometimes taken to be.

1. Introduction: **The "Replication Crisis"**

In the current debate about replicability in psychology, we can distinguish between (1) the question of why not more replication studies are done (e.g., Romero 2017) and (2) the question of why a significant portion (more than 60%) of studies, when they *are* done, fail to replicate (I take this number from the Open Science Collaboration, 2015). Debates about these questions have been dominated by two assumptions, namely, first, that it is in general desirable that scientists conduct replication studies that come as close as possible to the original, and second, that the low replication rate can often be attributed to statistical problems with many initial studies, sometimes referred to as "p-hacking" and "data-massaging."[1]

[1] An important player in this regard is the statistician Andrew Gelman who has been using his blog as a public platform to debate methodological problems with mainstream social psychology (http://andrewgelman.com/).

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

I do not wish to question that close (or "direct") replications can sometimes be epistemically fruitful. Nor do I wish to question the finding that are severe problems in the statistical analyses of many psychological experiments. However, I contend that the focus on formal problems in data analyses has come at the expense of questions about the notion of *replication* as such. In this paper I hope to remedy this situation, highlighting in particular the implications of the fact that psychological experiments in general are infused with conceptual and material presuppositions. I will argue that once we gain a better understanding of what this entails with respect to replication, we get a deeper appreciation of philosophical issues that arise in the investigative practices of psychology. Among other things, I will show that replication is not as central to these practices as it is often made out to be.

The paper has three parts. In part 1 I will briefly review some philosophical arguments as to why there can be no exact replications and, hence, why attempts to replicate always involve individuation judgments. Part 2 will address a distinction that is currently being debated in the literature, i.e., that between direct and conceptual replication, highlighting problems and limitations of both. Part 3, finally, will argue that a significant part of experimental research in psychology is geared toward exploring the shape of specific phenomena or effects, and that the type of experimentation we encounter there is not well described as either direct or conceptual replication.

2. The Replication Crisis and the Ineliminability of Concepts

When scientists and philosophers talk about successfully replicating an experiment, they typically mean that they performed the same experimental operations/interventions. But what does it mean to perform "the same" operations as the ones performed by a previous experiment? With regard to this question, I take it to be trivially true that two experiments cannot be identical: At the very least, the time variable will differ. Replication can therefore at best aim for *similarity* (Shavit & Ellison 2017), as is also recognized by some authors in psychology. In this vein, Lynch et al (2015) write that "[e]xact

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

replication is impossible" (Lynch et al 2015, 2), arguing that at most advocates of direct replication can aim for is to get "as close as possible," i.e., to conduct an experiment that is similar to the previous one. In the literature, such experiments are also referred to as "direct replications." (e.g., Pashler & Harris 2912).[2]

The notion of similarity is, of course, also notoriously problematic (e.g., Goodman 1955), since any assertion of similarity between A and B has to specify with regard to what they are similar. In the context of experimentation, the relevant kinds of specifications already presuppose conceptual and material assumptions, many of which are not explicated, about the kinds of factors one is going to treat as relevant to the subject matter (see also Collins 1985, chapter 2). Such conceptual decisions will inform what one takes to be the "experimental result" down the line (Feest 2016). For example, If I am interested in whether listening to Mozart has a positive effect on children's IQ, I will design an experiment, which involves a piece by Mozart as the independent variable and the result of a standardized IQ-test at a later point. Now if I get an effect, and if I call it a Mozart effect, I am thereby assuming that the piece of music I used was causally responsible *qua being a piece by Mozart*. Moreover, when I claim that it's an effect on intelligence, I am assuming that the test I used at the end of the experiment *in fact measured intelligence*. These judgments rely on conceptual assumptions already built into the experiment qua choice of independent and dependent variables. In addition, I need *material assumptions* to the effect that potentially confounding variables have been controlled for. I take this example to show that whenever we investigate an effect *under a description*, we cannot avoid making conceptual assumptions when determining whether an experiment has succeeded or failed. This goes for original experiments as well as for replications.

[2] Both advocates and critics of direct replication sometimes contrast such replications with "conceptual" replications" (Lynch et al 2015). We will return to this distinction below.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

One obvious rejoinder to this claim might be to say that replication attempts need not investigate effects under a description. They might simply imitate what the original experiment did, with no particular commitment to what is being manipulated or measured. But even if direct replications need not explicitly replicate effects under a description, I argue that they nonetheless have to make what Lena Soler calls "individuation judgments" (Soler 2011). For example, the judgment that experiment 2 is relevantly similar to experiment 1 involves the judgment that experiment 2 does not introduce any confounding factors that were absent in experiment 1. However, such judgments have to rely on some assumptions about what is relevant and what is irrelevant to the experiment, where these assumptions are often unstated auxiliaries. For example, I may (correctly or incorrectly) tacitly assume that temperature in the lab is irrelevant and hence ignore this variable in my replication attempt.

It is important to recognize that the individuation judgments made in experiments have a high degree of epistemic uncertainty. Specifically, I want to highlight what I call the problem of "conceptual scope," which arises from the question of how the respective independent and dependent variables are described. Take, for example, the above case where I play a specific piece by Mozart in a major key at a fast pace. A lot hangs on what I take to be the relevant feature of this stimulus: the fact that it's a piece by Mozart, the fact that it's in a major key, the fact that it's fast? etc. Depending on how I describe the stimulus, I might have different intuitions about possible confounders to pay attention to. For example, if I take the fact that a piece is by Mozart as the relevant feature of the independent variable, I might control for familiarity with Mozart. If I take the relevant feature to be the key, I might control for mood. Crucially, even though scientists make decisions on the basis of (implicit or explicit) assumptions about conceptual scope, their epistemic situation is typically such that they don't know what is the "correct" scope. This highlights a feature of psychological experiments that is rarely discussed in the literature about the replication crisis, i.e., the deep epistemic uncertainty and conceptual openness of much

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

research. This concerns both the initial and the replication study. Thus, concepts are ineliminable in experimental research, while at the same time being highly indeterminate.

3. Is the dichotomy between direct and less direct replication pragmatically useful?

One way of paraphrasing what was said above is that all experiments involve individuation judgments and that this concerns both original and replication studies. While this serves as a warning against a naïve reliance on direct (qua non-conceptual) replication, it might be objected that direct replications nonetheless make unique epistemic contributions. This is indeed claimed by advocates of both direct and less direct (="conceptual") replication alike. I will now evaluate claims that have aligned the distinction between direct and "conceptual" with some relevant distinctions in scientific practice, such as that between the aim of establishing the existence of a phenomenon and that of generalizing from such an existence claim on the one and that between reliability and validity on the other. I will argue that while these distinctions are heuristically useful, but on closer inspection bring to the fore exactly the epistemological issues just discussed.

3.1 Existence vs. Generalizability

Many scientists take it as given that there cannot be two identical experiments, but nonetheless argue that there is significant epistemic merit in trying to get *close enough*., i.e., to conduct direct replications. In turn, the notion of a direct replication is frequently contrasted with that of a "conceptual" replication. In a nutshell, direct replications essentially try to redo "the same" experiment (or at least something very close), whereas the conceptual replications try to operationalize the same question or concept/effect in a different way. The advantage of direct replications, as viewed by its advocates, is that by being able to redo an experiment faithfully and to create the same effect, one can show that the effect was real: "Exact and very close replications establish the basic existence and stability of a

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

phenomenon by falsifying the (null) hypothesis that observations simply reflect random noise" (LeBel et al, forthcoming, 7).

Advocates of conceptual replication don't deny this advantage of close replications, but hold that we want more than to establish that a given effect – created under very specific experimental conditions – is real. We want to know whether our findings about it can be generalized to: "When the goal is generalization, we argue that 'imperfect' conceptual replications that stretch the domain of research may be more useful" (Lynch et al 2015, 2). From a strictly Popperian perspective, the idea that non-falsification of the hypothesis of random error can provide proof of stability and existence is questionable , of course. But even if we abandon Popperian ideology here and take the falsification of H0 (that the initial effect was due to random error) to point to the truth of H1 (that there is a stable effect), the question is how to describe the effect. In other words, when claiming to have confirmed an effect, we have to say *what kind of effect* it is. And there we face the following dilemma:

a) Either we describe the effect as highly specific to very local experimental circumstances, involving the choice of a specific independent variable, delivered in a specific way etc.
b) Or we describe it in slightly broader terms, e.g., as a Mozart effect.

Advocates of direct replication might indeed endorse something like a), thereby exhibiting the kind of caution that motivated early operationists, in that no claim is made beyond the confines of a specific experiment. If, on the other hand, psychologists endorsed a description such as b), they would immediately run into the question of conceptual scope, i.e., the question *under what description* the independent variable can be said to have caused an effect. I argue that no amount of direct replication can answer this question, and hence, even if direct replication can confirm the existence of an effect, it cannot say what kind of effect. By asserting this, I am not saying that it's never useful to do a direct replication. My claim is merely that it will tell us relatively little. More pointedly: Direct replication can (perhaps) provide evidence for the existence of something, but it cannot say *existence of what*. Rolf

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

Zwaan makes a similar point when he states that "replication studies "tell us about the reliability of those findings. They don't tell us much about their validity." (Zwaan 2013).

In a similar vein, I argue that direct replication, with its narrow focus on ruling out random error, is epistemically unproductive, because it has nothing to say about *systematic error*. Systematic error arises if one erroneously attributes an effect to a specific feature of the experiment, when it is in fact due to another feature of the experiment. This can include, but is not limited to, the above-mentioned problem of conceptual scope. Fiedler et al. (2012) make a similar point when they argue that a narrow focus on falsification (with the aim of avoiding false positives) can be detrimental to the research process. Differently put, by privileging direct replication, we are not in a position to inquire about the kind of effect in question. This question, I argue, is best addressed by paying close attention to the possibility of systematic error, and hence by doing conceptual work. In other words, experimentally probing into systematic errors of conceptual scope is a valuable and productive part of the research process as it enables scientists to gradually explore what kind of effect (if any) they are looking at.[3]

3.2 Generality

I have argued that (a) scientists typically produce effects under a description and (b) that it can be epistemically productive to probe the scope of the description and to investigate the possibility of systematic error with regard to experiments that draw on such descriptions. It is epistemically productive, because it forces scientists to explore the nature and boundaries of the effect they are investigating. With this I have argued against a narrow focus on direct replication and I have cautioned against overstating the epistemic merits of such replication. But when we are concerned with effects

[3] I take this to be a contribution to arguments that philosophers of experimentation have made for a long time; e.g., Mayo 1996.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

under a description, we are confronted with questions about the adequacy of the description. It is this question that advocates of "conceptual replication" claim to be able to address when they emphasize that their approach can deliver generality (over mere existence).

We have to distinguish between two notions of generality, namely (a) what kinds of descriptions one can generalize or infer to *within the experiment*, and (b) does the effect in question hold *outside the lab* (see Feest & Steinle 2016). These types of generality are also sometimes referred to as internal vs. external validity, respectively (Campbell & Stanley 1966; Guala 2012), where the former refers to the quality of inferences within an experiment and the latter refers to the quality of inferences from a lab to the world. The notion of generalizability raises questions about two kinds of validity. My focus here will be on internal validity, i.e., with the question of whether the effect generated in an experiment really exists as described by the scientist.[4]

Internal validity can fail to hold because of epistemic uncertainties regarding confounding variables both internal and external to experimental subjects. For example, prior musical training might make a difference to how one responds to Mozart music, but the experimenter may not have taken this into consideration in their design. But internal validity can also fail to hold is by virtue of what I have referred to as the problem of conceptual scope (for example, we may refer to the effect as a Mozart effect when it is in fact a Major-key effect). Effectively, when I treat a major-key effect as a Mozart effect, I have misidentified the relevant causal feature of the stimulus. In turn, this means that I will neglect to control for major/minor key as I will regard this as irrelevant, which can result in systematic errors. In both cases, scientists can go wrong in their individuation judgment. What is at stake is not whether there is an effect, but what kind of effect it is. Now, given that those kinds of problems can

[4] In this respect I differ from some advocates of conceptual replication, who have highlighted external validity as a desideratum (E.g., Lynch 1982, 3/4).

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

occur, we turn to the question of whether "conceptual replication" has an answer. I will now argue that it does not.

To explain this, let me return to the above characterization of conceptual replication, according to which such replication consists in repeating an experiment, using different operationalizations of the same construct. For example, a conceptual replication of an experiment about the Mozart effect might operationalize the concept Mozart effect differently by using a different piece of Mozart music and/or a different measure of spatial reasoning. But there is a major caveat here: If I want to compare the results of two experiments that operationalized the same construct differently, I already have to presuppose that both operationalizations in fact have the same conceptual scope, i.e., that they in fact individuate the same effect. But this would be begging the question, since after all – given the epistemic uncertainty and conceptual openness highlighted above – that's precisely what's at issue. Differently put, experiment 2 might or might not achieve the same result as experiment 1, but the reason for this would be underdetermined by the experimental data. Thus, the problem of conceptual scope prevents us from being able to say whether we have succeeded in our conceptual replication.

Given the uncertainties as to whether one has in fact succeeded in conceptually replicating a given experiment, I am weary of the language of replication here. If anything, I would argue that the method in question should be regarded as a research strategy that is aimed at helping to demarcate and explore the very subject matter under investigation. But as I will argue now, this is perhaps better described as exploration, not as replication.

4. Putting Replication in its Proper Place

The conclusion of the previous paragraphs seems pretty bleak: Direct replication is either extremely narrow in what it can deliver or it runs into the joint problems of confounders and conceptual scope. Conceptual replication, on the other hand, cannot come to the rescue, because it also runs into the

exact same problems. Should we then throw up our hands and conclude that since ultimately neither direct nor conceptual replication are possible the crisis of replication is much more severe than we previously thought? This would be the wrong conclusion, however. This would only follow if replication was in fact as central to research as it is sometimes taken to be. I claim that it is not. My argument for these claims has three parts. The first part holds that exploring (the possibility of) systematic errors is an important part of the investigative process, which is not well described as replication. Second, if we take seriously this process of exploring and delineating the relevant phenomena, we find that there is indeed a great deal of uncertainty in psychological research, but this, in and of itself, does not necessarily constitute a crisis. Lastly, while it is fair to say that there is a crisis of confidence in current psychology, it is not well described as a replication crisis.

Let me begin with the first point. I have argued that direct replication (even where it is successful) is of limited value, because it can at most rule out random error, but completely fails to be able to address systematic error. But if we appreciate (as I have argued we should) that direct replication inevitably involves individuation judgments, it is obvious that there is always a danger of systematic error, because I have to assume that all confounding variables have been controlled for. One important class of confounders follows from what I have referred to as the problem of conceptual scope, i.e., the difficulty of correctly describing both the independent variable responsible for a given effect and the dependent variable.[5] Epistemically productive experimental work, I claim, therefore needs to focus on systematic errors, specifically those brought about by unstated auxiliary assumptions.

Indeed, if we look at the story of the Mozart effect, we find that this is exactly what happened. This example also nicely illustrates my claim about the conceptual openness and epistemic uncertainty

[5] My focus here has been mainly on the former. But of course the problem of conceptual scope concerns both.

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

in many areas of experimental psychology. The Mozart effect was first posited by Rauscher and colleagues (Rauscher et a. 1993). It can now be regarded as largely debunked. While it is true that several people tried (and failed) to replicate the effect (e.g., Newman et al. 1995; Steele 1997), it is important to look at the details here. It is not the case that the effect was simply abandoned for lack of replicability. Rather, when we look at the back and forth between Rauscher and her critics, we find that the discussion turned on the choices and interpretations of independent and dependent variables. In this vein, Newman et al (1995) and Steele (1997) used different dependent variables, prompting Rauscher to argue that her effect was more narrowly confined to the kind of spatial reasoning measured by the Stanford-Binet. I suggest that we interpret this case as one where Rauscher was forced to confront (and retract) an unstated auxiliary assumption of her initial study, namely that the spatial reasoning subtest of the Stanford-Binet (which she had used as her dependent variable), was representative of spatial reasoning more generally. Likewise, her choice of the Mozart's Sonata for Two Pianos in D-major as the independent variable was put under considerable pressure by critics, who suggested that the relevant feature of the independent variable was not that it was a piece by Mozart, but that it was up-beat and put subjects in a good mood (Chabris 1999). My point here is that the debates surrounding the Mozart effect are best described as conceptual work, exploring consequences of possible errors that might have arisen from the problem of conceptual scope. At issue, I claim, was not primarily whether Rauscher really found an effect, but rather what was the scope of the effect.

I argue that this is a typical case. Rather than, or in addition to, attempting to conduct direct replications of previous experiments, researchers critically probed some hidden assumptions built into the design and interpretation of the initial experiment. My point here is both descriptive and normative. Thus, I argue that this is a productive way to proceed. However, I claim that it is not well described as replication, let alone conceptual replication. Rather, what we see here is a case in which scientists explore the empirical contours of a purported effect in the face of a high degree of epistemic

uncertainty and conceptual openness, and this is precisely why the case is not well described as employing conceptual replication. The reason for this is quite simple: For a conceptual replication to occur, one needs to already be in the possession of some well-formed concepts, such that they can be operationalized in different ways. It also presupposes that in general the domain is well-understood, such that operationalizations can be implemented and confounding variables can be controlled. But this completely misses the point that researchers often investigate effects precisely because they don't have a good understanding (and hence concept) of what it is.

Therefore I argue that while direct replication can only contribute a very small part to the research process, conceptual replication cannot make up for the shortcomings of direct replication. Instead, productive research should (and frequently does) proceed by exploring, and experimentally testing, hypotheses about possible systematic errors in experiment. Such research, I suggest, can contribute to conceptual development by helping to explore and fine-tune the shape and scope of proposed or existing concepts. The fact that this is riddled with problems does not in and of itself constitute a crisis, let alone a replication crisis.

5. Conclusion

The upshot of the above is that when we talk about the importance of replication, we need to be clear on what we mean by replication and why it is so important, precisely.

In this paper I have argued that if by replication we mean either "direct" or "conceptual" replication, we need to first of all be clear that direct replications are not non-conceptual. I then turned to some alleged epistemic merits of direct replication, for example that they can establish the existence of effects or the reliability of procedures that detect effects. I argued that insofar as such replications involve concepts, they run (among other things) into the problem of conceptual scope, i.e., the difficulty of determining, on the basis of independent and dependent variables of experiments what precisely is

Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

the scope of the effect one is trying to replicate. I highlighted that this is a real and pernicious problem in experimental research in psychology, due to the high degree of epistemic uncertainty and conceptual openness of many fields of research.

While my emphasis of the conceptual nature of replication may suggest that I would be more favorably inclined toward conceptual replication, I have argued that conceptual replication runs into the same problems, and for similar reasons: The very judgement that one has successfully performed a conceptual replication of a previous experiment presupposes what is ultimately the aim of the research, namely to arrive at a robust understanding of the relevant area of research. This, I argue that since conceptual replication presupposes a relatively good grasp of the relevant concepts, it is begging the question, and I suggested instead that researchers (should) engage in a process of specifically investigating possible systematic errors in original studies as a means to develop the relevant concepts. This process is not best described as one of replication, however. Summing up, then, I conclude that in general, replications are less useful and important than is widely assumed – at least in the kind of psychological research I have focused on in this paper.

Now, in conclusion let me return to the notion of a crisis in psychology as it is currently discussed in the literature. Obviously, I do not mean to deny that there is a crisis of confidence in (social) psychology (Earp & Trafimov 2015) as well as in other areas of study. However, based on the analysis provided in this paper, I argue that this crisis is not well described as a crisis of replication. Rather, it seems to be to a large degree a crisis that turns on questionable research practices with regard to the use of statistical methods in psychology (see Gelman & Loken 2014). While acknowledging the valuable philosophical and scientific work that is being done in this area, I suggest that a broader focus on the notion of replication provides us with a deeper appreciation of the conceptual dynamics characteristic of experimental practice.

REFERENCES

Campbell, D. T., and Stanley, J. C. (1966), Experimental and Quasi-Experimental Designs for Research (Chicago: Rand McNally).

Chabris, C. (1999): Prelude or requiem for the 'Mozart Effect?' "Scientific Correspondence", Nature, 4000, 826.

Collins, H. (1985). Changing order. Replication and induction in scientific practice. Chicago and London: The University of Chicago Press.

Earp, Brian & Trafimow, David (2015): "Replication, falsification, and the crisis of confidence in social psychology." Front. Psychol, 19 May 2015 | https://doi.org/10.3389/fpsyg.2015.00621

Feest, U., 2016, "The Experimenters' Regress Reconsidered: Tacit Knowledge, Skepticism, and the Dynamics of Knowledge Generation". Studies in History and Philosophy of Science, Part A 58 34-45.

Feest, U. & Steinle, F., 2016, "Experiment." In P. Humphreys (Ed.): Oxford Handbook of Philosophy of Science. Oxford University Press, 274–295.

Fiedler, K.; Kutzner, F. & Krueger, J. (2012): „The Long Way from alpha-error control to validity proper: Problems with a short-sighted false-positive debate." Perspectives on Psychological Science 7(6), 661-669

Gelman, Andrew & Loken, Eric (2014): The Statistical Crisis in Science. Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up. American Scientist 102 (6) 460-464. DOI: 10.1511/2014.111.460

Goodman, Nelson (1983/1955): Fact. Fiction and Forecast. Harvard University Press; 4 Revised edition edition

Guala, F. (2012), "Philosophy of Experimental Economics." In U. Mäki (ed.), Handbook of the philosophy of science. Vol. 13: Philosophy of Economics (Boston: Elsevier/Academic Press), 597–640

LeBel, E.P.; Berger, D., Campbell, L.; Loving, T. (2017): "Falisifiability is not Optional." Journal of Personality and Social Psychology (forthcoming)

Lynch, J. (1982): "On the External Validity of Experiments in Consumer Research. Journal of Consumer Research 9, 225-239. (December)

Lynch, J.; Bradlow, E.; Huber, J.; Lehmann, D. (2015): "Reflections on the replication corner: In praise of conceptual replication." IJRM ???

Mayo, Deborah (1996): Error and the Growth of Experimental Knowledge. University of Chicago Press.

Newman, J., Rosenbach, J., Burns, K.; Latimer, B., Matocha, H., Vogt, E. (1995: An experimental test of the 'Mozart Effect': Does listening to Mozart improve spatial ability? Perceptual and Motor Skills, 81, 1379-1387.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349

Pashler, Harold & Harris, Christine (2012): "Is the Replication Crisis Overblown?" Perspectives on Psychological Science 7(6), 531-536.

Rauscher, F., Shaw, G.; Ky, K. (1993). Music and spatial task performance. Nature ,365, 611.

Romero, Felipe (2017): "Novelty vs. Replicability. Virtues and Vices in the Reward System of Science." Philosophy of Science.

Shavit, Ayelet & Ellison, Aaron (eds.) (2017): Stepping in the Same River Twice. Replication in Biological Research. Yale University Press

Soler, Lena (2011): "Tacit Elements of Experimental Practices: analytical tools and epistemological consequences." European Journal for Philosophy of Science 1, 393-433.

Steele, K., (2000). Arousal and mood factors in the 'Mozart effect'. Perceptual and Motor Skills, 91, 188-190.

Zwaan, Rolf (2013): "How Valid are our Replication Attempts?"

https://rolfzwaan.blogspot.de/2013/06/how-valid-are-our-replication-attempts.html

Author: Paul L. Franco, UW-Seattle, Department of Philosophy

Contact: pfranco@uw.edu

Title: Speech Act Theory and the Multiple Aims of Science

Abstract: I draw upon speech act theory to understand the speech acts appropriate to the multiple aims of scientific practice and the role of nonepistemic values in evaluating speech acts made relative to those aims. First, I look at work that distinguishes explaining from describing within scientific practices. I then argue speech act theory provides a framework to make sense of how explaining, describing, and other acts have different felicity conditions. Finally, I argue that if explaining aims to convey understanding to particular audiences rather than describe literally across all contexts, then evaluating explanatory acts directed to the public or policymakers involves asking nonepistemic questions.

(Accepted with minor revisions to the PSA 2018 proceedings issue of Philosophy of Science | Revisions not yet made; final version due January 2019)

1. Introduction

Hasok Chang "[complains] about…our [i.e., philosophers of science] habit of focusing on descriptive statements that are either products or presuppositions of scientific work, and our commitment to solving problems by investigating the logical relationships between these statements" (2014, 67–8). He argues philosophers of science should adopt "a change of focus from propositions to actions" (67). Chang suggests, "When we do pay attention to words, it would be better to remember to think of 'how to do things with words', to recall J. L. Austin's (1962) famous phrase" (68).

In this paper, I take Chang's suggestion and argue that attending to Austin's account of the things we do with words can help us understand the multiple goals of scientific practices, the speech acts appropriate to those goals, and the roles of nonepistemic values in evaluating speech acts made relative to those aims. In §2, I give an overview of a few philosophers of science working on explanation who have shifted focus from propositions to explaining.[1] I also briefly relate this work to a few themes in speech act theory. In §3, I give more details of Austin's framework to highlight ways of evaluating speech acts beyond truth and falsity. In §4, I explore the multiple goals of scientific practice, especially goals related to conveying understanding to the general public and policymakers, and the speech acts appropriate to those goals.

2. The things scientists do with words

2.1 Explaining

Consider some recent and not-so-recent work on scientific explanation. Andrea Woody's defense of a functional perspective on explanation aims to motivate "a shift in focus away from explanations, as achievements, toward explaining, as a coordinated activity of communities" (2015, 80). In a similar spirit, Angela Potochnik argues that when looking at explanation, "sidelining the communicative purposes to which explanations are put is a mistake" (2016, 724). She emphasizes that explaining is a communicative act involving a speaker and audience made against a background that shapes the explanations offered. In so

[1] I make no claims Chang influenced the work I canvas.

arguing, Potochnik deliberately recalls Peter Achinstein's claim, "Explaining is an illocutionary act," i.e., a speech act uttered by a speaker with a certain force and for a certain point (1977, 1).

These accounts share in common an emphasis on the importance of the aims of the speaker and audience, and thus the context of utterance in evaluating, to borrow terminology from Austin, the felicity conditions of explanatory speech acts. In particular, we might focus on the aims of the speaker and their audience in requesting and giving explanations, the time and location of an explaining speech act, and, following Woody, "what role(s) [explanations] might play in practice" (2015, 81). In focusing on the explaining act rather than the supposedly stable propositional content of an act of explanation, our attention is drawn to dimensions of evaluation beyond truth and falsity.

On this last point, Nancy Cartwright argues that the functions of a scientific theory to "tell us…what is true in nature, and how we are to explain it…are entirely different functions" (1980, 159). *Ceteris paribus* laws used in scientific theories are literally false, but still do explanatory work. One way to understand Cartwright's claim is that the speech act of describing the world truly and the speech act of explaining come apart from one another. In coming apart from one another and fulfilling different aims within scientific practice, descriptive and explanatory speech acts have different felicity conditions. For example, Potochnik (2016) examines the ways in which explaining increases understanding. But, Potochnik argues, what gets explained depends on a speaker's and audience's interests, and an explaining act's success in generating understanding depends on the cognitive resources of the audience. As such, to evaluate any given communicative act of explaining requires attending to the epistemic and nonepistemic interests of speakers and audiences that form the background against which explanations are offered. This means evaluating explanatory speech acts solely in terms of truth or falsity is inapt.

2.2 Multiple aims and the true/false fetish

I do not think this focus on acts and away from the truth or falsity of descriptive statements is unique to philosophers of science interested in explanation. We see a similar shift in work on the so-called aims approach to values in science (e.g., Elliott and McKaughan 2014;

Intemann 2015). The aims approach shares in common with work on explaining a recognition that scientific practice aims at more than describing the world truly or falsely. Further, if some of those aims include things like making timely policy recommendations for decision makers or increasing public understanding of science, there is a role for nonepistemic values in parts of scientific practice. As Kevin Elliott and Daniel McKaughan put this point, "representations can be evaluated not only on the basis of the relations that they bear to the world but also in connection with the various uses to which they are put" (2014, 3).

Why look to speech act theory to flesh out this picture about the multiple aims of scientific practice and their relationship to nonepistemic values? In part because speech act theory makes sense of the different uses to which one and the same sentence might be put depending on the aims of the speaker and audience and the context of utterance. In doing so, I think Austin is right that we can "play Old Harry with two fetishes…(1) the true/false fetish, (2) the value/fact fetish" (1962, 150). Austin was mainly content to play Old Harry with these fetishes to free philosophers from the grip of the so-called descriptive fallacy: the view "that the sole business, the sole interesting business, of any utterance…is to be true or at least false" (1970, 233). But I also think that in combating the descriptive fallacy and the true/false and fact/value fetishes, speech act theory motivates a constructive shift from the truth or falsity of descriptive statements to the things we do with words.

Take Austin's claim that evaluating apparently descriptive speech acts like "'France is hexagonal,'" involves nonepistemic questions about who is uttering the statement, in what context, and with what "intents and purposes" (1962, 142). Rather than concluding the sentence is false and leaving it at that, Austin points out the different speech acts one can use such a sentence to perform, e.g., stating or interpreting or estimating. In determining the use the sentence is put to—with the help of context and by inquiring after the interests of the speaker and their audience—we might realize, irrespective of the sentence's literal truth or falsity, "It is good enough for a top-ranking general, perhaps, but not for a geographer" (142). In other words, it serves the aims of the general, which, unlike the aims of the geographer, do not necessarily require a descriptively literal account of France's shape. The statement might not aim to assert or describe literally, but do something else entirely. As such,

evaluating it along the lines of truth or falsity will miss something important about the aims of a speaker in uttering it.

To expand on this picture, I turn to explicating Austin's speech act theory.

3. Austin's speech act theory

3.1 Performatives and constatives

Austin first drew our attention to the things we do with words by discussing performative utterances. Austin says of these, "if a person makes an utterance of this sort we should say that he is *doing* something rather than merely *saying* something" (1970, 235). Imagine a speaker utters 'I promise to return my referee report in two weeks' during the peer review process. In making this speech act, Austin claims the speaker does not describe an internal act she has concurrent to her utterance. Instead, in making that utterance, the speaker just is performing the act of promising thereby committing herself to actions related to the timely review of papers.

While promising has no special connection to truth and falsity, it still must meet what Austin calls felicity conditions to be happy or unhappy. In order to promise to return their referee report in two weeks successfully, the speaker must meet the sincerity condition of forming an intention to do so, even if they are not describing "some inward spiritual act of promising" (236). The speaker must also be in a position to follow through on their intention. Thus, there is unhappiness in the speech act if the speaker promises knowing full well other commitments will prevent her from returning the report in two weeks. The speaker must also have the authority to make a promise; unless authorized, an editor cannot promise on behalf of a reviewer. There should also exist a convention for making a promise in peer review contexts. Such conventions might allow the speaker to promise without uttering, 'I promise,' e.g., by accepting a request that reads, 'In accepting this review assignment you commit to returning the referee report within such-and-such a time.'

Austin first contrasts performatives with constatives, e.g., descriptive statements or assertions that aim to state something truly or falsely about the world, but which do not seem to perform an action. However, Austin claims describing or asserting is as much an action as promising, even if the felicity conditions for asserting are more closely connected to truth

or falsity. Consider an editor saying of a reviewer, 'They review quickly, and I expect that they will return their review within two weeks.' In saying this, the editor commits herself to providing evidence for her description of the reviewer as quick, and perhaps justifying her expectation that the reviewer's past behavior provides good evidence for future behavior. As Robert Brandom puts this point, "In asserting a claim one not only authorizes further assertions, but commits oneself to vindicate the original claim, showing that one is entitled to make it" (1983, 641). That is, the utterer must be in a position of authority—here in an epistemic sense—with regards to the claim and be ready to perform further speech acts if so prompted. Other felicity conditions of assertions or descriptions include a sincerity condition: an editor uttering our example sentence should believe what they say. Finally, the context of an assertion also shapes its felicity conditions: an editor should utter the sentence in the appropriate circumstances, e.g., as a response to a worry about the speed of the review process. Should these conditions not be met, the speech act might be unhappy even if true.

3.2 Locution and illocution

Austin develops speech act theory to capture the similarities between performatives and constatives. Speech acts like promising and describing have three dimensions: the locutionary content, which is the conventional sense and reference of the uttered sentence; the illocutionary force, which is the use the utterance is put to; and the perlocutionary effects, which are intended and unintended "effects upon the feelings, thoughts, or actions of the audience, or of the speaker, or of other persons" (1962, 101).

Austin's points about the illocutionary dimension of a speech act most clearly capture how one and the same representation might be put to different uses depending on our goals, and how different uses have different felicity conditions despite sharing locutionary content. Consider the sentence, 'This product contains chemicals known to the state of California to cause cancer.' The locutionary content would just consist in the proposition expressed by the sentence as determined by the conventional sense and reference of the words. This content can be common to different illocutionary acts. Someone uttering the sentence could be describing a product, issuing a warning, or explaining why they do not use this particular product but another. Uttering the sentence with the force of a description, the force of a

warning, and the force of an explanation will have similar felicity conditions related to truth and falsity. Namely, the locutionary content should be true or approximately true for an utterance to count as a good description, a good warning, or a good explanation.

However, a warning might be infelicitous in ways a description might not. For example, warnings might be issued only in the case in which some pre-determined level of significant risk at a certain level of exposure is met. In cases where such levels are not met, issuing a warning might be infelicitous. Consider also that uttering such a sentence with the force of an explanation might be called for only if, e.g., someone is prompted to justify their choice of a product that does not contain cancer-causing chemicals over a more easily available and cheaper product that does contain those chemicals. In these last two cases, nonepistemic reasons related to risk, cost-effectiveness, and so on can enter into the evaluation of the happiness of a warning or explanation.[2]

Austin thinks attending to these points combats a form of abstraction that distorts our thinking about the felicity conditions of descriptive statements. He thinks that when examining statements, "we abstract from the illocutionary…aspects of the speech act, and we concentrate on the locutionary" (1962, 144–5). In so doing, "we use an over-simplified notion of correspondence with the facts—over-simplified because essentially it brings in the illocutionary aspect" (145). Such an approach focuses on "the ideal of what would be right to say in all circumstances, for any purpose, to any audience, &c." (145). But, as Austin claims, questions concerning correspondence with the facts brings with it the illocutionary aspect since truth or falsity does not attach to sentences or locutionary content. Instead, truth or falsity is related to particular things speakers do with sentences. Descriptions might be, strictly speaking, true or false, but not recommendations or explanations. In order to know, then, if evaluating a speech act along the true-false dimension is apt, we need to know the illocutionary force of that act. But to know the illocutionary force of the act requires we attend to context, including the aims of both speaker and audience, time and place of utterance, and conventions governing the specific speech situation. In this way, Austin

[2] Any speech act will also have perlocutionary effects, and we might follow Heather Douglas (2009) and Paul Franco (2017) in focusing on the nonepistemic consequences of making false descriptions, giving bad warnings, or explaining unclearly.

argues context and aims are central to determining the illocutionary force of a speech act, and hence to evaluating its felicity or infelicity.

4. Aims-approaches and speech act theory

4.1 Explaining and understanding

Scientific practice might seem to deal in paradigmatically constative speech acts, e.g., descriptions. Such speech acts are, to varying degrees, evaluable along dimensions of truth or falsity in ways we might question the relevance of speech act theory to philosophy of science. That is, we might say that scientific practice just is a case in which abstracting away from the illocutionary force of an utterance to focus on locutionary content is appropriate. For example, Austin says that "perhaps with mathematical formulas in physics books…we approximate in real life to finding" speech acts where focusing on the locutionary content is appropriate (1962, 145). If scientific practice aims at timeless truths holding across all contexts independent of the sorts of aims and interests of speakers and audiences necessary to evaluating the felicity or infelicity of speech acts, then it seems speech act theory is irrelevant to philosophy of science.

Yet, as Austin points out, "When a constative is confronted with facts, we in fact appraise it in ways involving the employment of a vast array of terms which overlap with those that we use in the appraisal of performatives. In real life, as opposed to the simple situations envisaged in logical theory, one cannot always answer in a simple manner whether it is true or false" (141–2). Consider again 'France is hexagonal.' Austin asks, "How can one answer…whether it is true or false that France is hexagonal? It is just rough, and that is the right and final answer to the question of the relation of 'France is hexagonal' to France. It is a rough description; it is not a true or false one" (142). Though rough, it is still open to evaluation. We can ask if it is in accord with conventions governing estimations and if this estimation serves the purposes and interests of the speaker and their audience at the time of utterance. 'France is hexagonal' can count as felicitous even if rough and not literally true because it aims at something other than truth.

Austin claims that many of our apparently constative speech acts are evaluable along similar dimensions given that they also confront facts in similarly rough ways. McKaughan

makes a related point about scientific speech acts. He argues that certain speech acts central to scientific practice like "conjecturing, hypothesizing, guessing and the like often play a role in scientific discourse that serves neither to assert that an hypothesis is true nor to express such a belief" (2012, 89). Moreover, as mentioned in §2, the picture of scientific practice as concerned solely with the truth is challenged, among other places, in work on explanation, and also in values in science. For example, when looking at the role particular acts or patterns of explaining play in scientific discourse we might focus not on the locutionary content of an explanatory speech act, but on the ways "explanatory discourse…functions to sculpt and subsequently perpetuate communal norms of intelligibility" (Woody 2015, 81). In focusing on this aspect of explaining, we might find, for example, that "the ideal gas law's role in practice is not essentially descriptive, but rather prescriptive; by providing selective attention to, and simplified treatment of, certain gas properties (and their relations) and ignoring other aspects of actual gas phenomena, the ideal gas law effectively instructs chemists in how to think about gases as they are characterized within chemistry" (82). In other words, the ideal gas law, in practice, does not have the force of a descriptive speech act, but lays down a rule of sorts guiding the investigation of gases.[3] The success of acts of explaining from this perspective will have less to do with accurately describing actual gases, but the way they facilitate, say, the education of new scientists or increase understanding of related phenomena, e.g., "by laying foundation for the concept of 'temperature'" beyond "the subjective, inherently comparative quality of human perception" (82). An act of explaining that fails to achieve pedagogical aims or fails to increase understanding of related phenomena might be infelicitous even if the locutionary content of that act confronts the facts in the right way to count as approximately true.

On this point about the ways explanations might increase understanding without describing, Potochnik claims "that what best facilitates understanding is not determined solely by the relationship between a representation and the world" (2015, 74). An idealized explanation like the ideal gas law is not defective because it fails to fully describe all the

[3] About universal generalizations Austin writes, "many have claimed, with much justice, that utterances such as those beginning 'All…' are prescriptive definitions or advice to adopt a rule" (1962, 143). Austin does not fully endorse this suggestion.

possible causal factors at play in the behavior of actual gases. Though literally false, an idealization might be successful insofar as it "secure[s] computational tractability" or successfully isolates "all but the most significant causal influences on a phenomenon" (71). In so doing, we increase our understanding by facilitating "successful mastery, in some sense, of the target of understanding" or "by revealing patterns and enabling insights that would otherwise be inaccessible" (72). Indeed, pointing out all the ways in which the ideal gas law fails to hold for actual gases or is literally false as a description might hinder the use of explanations in scientific discourse to provide "shared exemplars that function as norms of intelligibility" (Woody 2015, 84).

In a related vein, Potochnik argues, "Because understanding is a cognitive state, its achievement depends in part on the characteristics of those who seek to understand," including both the speaker and the audience (2015, 74). In evaluating an act of explaining, we should look at how the speaker's interest has shaped the focus of their explanation and also how the explanation increases an audience's understanding, where this involves considering the audience's interests in seeking an explanation. An explanation that fails to be relevant to the audience or fails to increase their understanding or guide their thinking about related phenomena, but that nonetheless has locutionary content that is approximately true, might count as infelicitous.

4.2 Values and science

On the views of explaining canvassed, the aims of generating literally true descriptions of the world come apart from, say, explaining and understanding the most important causal factors at play for a given phenomenon. Now, as the aims approach to the proper role for nonepistemic values in scientific practice emphasizes, explaining and describing do not exhaust the goals of scientific practice. The aims approach focuses on the ways "scientific decision-making, including methodological choices, selection of data, and choice of theories or models, are...a function of the aims that constitute the research context" (Intemann 2015, 218). Given that the research context includes social, political, and moral considerations, the aims of science can just as well be understood in nonepistemic ways as it can be understood in epistemic ways.

Consider, for example, the American Geophysical Union's position statement on human-induced climate change. At the end of their statement, they claim, "The community of scientists has responsibilities to improve overall understanding of climate change and its impacts. Improvements will come from pursuing the research needed to understand climate change, working with stakeholders to identify relevant information, and conveying understanding clearly and accurately, both to decision makers and to the general public" (American Geophysical Union 2013). Here, I focus on the claim that scientists have responsibilities to improve the understanding of policymakers and the general public, and drawing upon the aforementioned work on explaining, think about how adopting this aim shapes the felicity conditions of explanatory speech acts directed at the audiences mentioned.

Notice that the position statement distinguishes the research necessary to understand climate change from conveying that understanding to policymakers and the general public. The sense in which these different activities come apart from one another and have different success conditions can be made sense of, in part, by focusing on the audience to whom scientists are speaking. We saw that for Potochnik (2016) understanding is a cognitive state that depends on the abilities and interests of those who are explaining and those to whom explanations are directed. In communicating to policymakers and the general public, scientists should consider the interests of the speaker in asking for an explanation as well as their level of knowledge regarding the phenomenon in question, in this case, climate change. In so doing, scientists might find that a description that aims to describe climate change in all its complexity might not serve these aims well. Instead, scientists might aim for an explanation that, though omitting descriptive complexity, draws upon models that represent those causal factors related to the audience's interests in a way that is cognitively accessible and helps guide the public in thinking more generally about climate change.

On this point, the American Geophysical Union's position statement maintains scientists ought to enlist the help of stakeholders in identifying potentially relevant information to their research. This is a point Intemann makes in developing the aims approach. She says of climate science, "[T]he aim is not only to produce accurate beliefs about the atmosphere, but to do so in a way that allows us to generate useful predictions for protecting a variety of social, economic and environmental goods that we care about" (2015,

219). In the view of the American Geophysical Union, in order to do this well, scientists ought to consult with relevant stakeholders and policymakers regarding what they value. Thus, for example, if stakeholders and policymakers communicate worries about extreme weather events and "how to adapt to 'worst case scenarios,' then models able to capture extreme weather events should be preferred" to those models that "anticipate slow gradual changes" (Intemann 2015, 220). Notice that in making such a decision, the grounds for choosing models able to represent aspects of climate change relevant to stakeholders' interests are nonepistemic rather than epistemic, e.g., generating predictions useful for protecting goods the general public cares about. Insofar as the representations or explanations generated do not meet these goals because they are unrelated to stakeholders' interests, the attendant speech acts might very well be infelicitous even if they describe some related phenomenon more or less accurately.

Both points about pitching explanations at a level that is cognitively accessible and choosing models for representing climate change phenomena in ways sensitive to stakeholders' interests illustrate a point Austin makes about the importance of uptake to successfully performing a speech act. Austin claims, "Unless a certain effect is achieved, the illocutionary act will not have been happily, successfully performed….I cannot be said to have warned an audience unless it hears what I say and takes what I say in a certain sense….Generally the effect amounts to bringing about the understanding of the meaning and force of the locution" (1962, 116). In aiming to convey understanding through explaining relevant aspects of climate change to decision makers and the general public, a speaker should consider the interests, background knowledge, and cognitive resources of their audience. Insofar as scientists fail to do so in explaining to the general public, even if the locutionary content that comprises their speech act approximates truth, they will not secure uptake in the sense of generating understanding in their audience. As such, their speech act will be infelicitous.

Of course, a scientist's explaining something to their audience will also be infelicitous if it is based on inaccurate information or extrapolates from what is known to their audience's interests in unjustified ways. However, this does not mean that if scientists aim to convey understanding to the public they should stick solely to descriptive claims. As

Elliott emphasizes in discussing how scientists should best communicate uncertainty to the public, "It does little good to expect scientists to provide unbiased information to the public if their pronouncements are completely misinterpreted or misused by those who receive them" (2017, 89). Similarly, "members of the public might not be able to 'connect the dots'" between scientists' descriptive speech acts and the ways those are relevant to their interests; insofar as scientists do not explain with the aims of conveying understanding—which as Potochnik argues, comes apart from describing the world truly in all its complexity—the public "would be left wondering what [the descriptions] might mean" (88). Thus, if scientists are to meet responsibilities the American Geophysical Union claims they have with regard to conveying understanding to the general public, those scientists should communicate using speech acts best able to secure uptake in the general public. This involves considering the interests and cognitive resources of the general public in ways that shape the felicity conditions of the speech acts beyond truth and falsity.

5. Conclusion

I argued speech act theory can tie together a few threads in recent work on explaining and values in science that share in common a shift in focus from descriptive propositions to things scientists do with words. Some of those things, like explaining, also seem the sorts of speech acts appropriate for fulfilling aims scientists have other than describing the world literally, like conveying understanding to the public and policymakers. Insofar as successfully fulfilling these aims involves explaining, and insofar as acts of explaining that secure uptake require attention to the nonepistemic interests and cognitive resources of speaker and audience, our attention is drawn towards ways explanatory speech acts can be happy or unhappy beyond describing truly or falsely. Future work will aim to delineate these felicity conditions in greater detail with an eye towards revealing further nonepistemic dimensions of evaluation.

References

Achinstein, Peter. 1977. "What is an Explanation?" *American Philosophical Quarterly* 14(1):1–15.

American Geophysical Union. 2013. "Human-Induced Climate Change Requires Urgent Action." https://sciencepolicy.agu.org/files/2013/07/AGU-Climate-Change-Position-Statement_August-2013.pdf

Austin, J.L. 1962. *How to Do Things With Words*. Ed. J.O. Urmson. Oxford: Oxford University Press.

----. 1970. "Performative Utterances." *Philosophical Papers*, 2nd edition. Eds. J.O. Urmson and G.J. Warnock. Oxford: Oxford University Press: 233–252.

Brandom, Robert. 1983. "Asserting", *Nous* 17(4):637–650.

Cartwright, Nancy. 1980. "The Truth Doesn't Explain Much." *American Philosophical Quarterly* 17(2):159–163.

Chang, Hasok. 2014. "Epistemic Activities and Systems of Practice: Units of Analysis in Philosophy of Science After the Practice Turn." *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science*, eds. Léna Soler, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost. New York: Routledge: 67–79.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, Kevin. 2017. *A Tapestry of Values*. New York: Oxford University Press.

Elliott, Kevin C. and Daniel J. McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81(1):1–21

Franco, Paul L. 2017. "Assertion, Nonepistemic Values, and Scientific Practice." *Philosophy of Science* 84(1):160–180.

Intemann, Kristen. "Distinguishing Between Legitimate and Illegitimate Values in Climate Modeling." *European Journal of the Philosophy of Science* 5:217–232.

McKaughan, Daniel J. 2012. "Speech acts, attitudes, and scientific practice: Can Searle handle 'Assuming for the sake of Hypothesis'?" *Pragmatics and Cognition* 20:1:88–106.

Potochnik, Angela. 2015. "The Diverse Aims of Science." *Studies in History and Philosophy of Science Part A* 53:71–80

----. 2016. "Scientific Explanation: Putting Communication First." *Philosophy of Science*, 83:721–732.

Woody, Andrea. 2015. "Re-orienting discussions of scientific explanation: A functional perspective." *Studies in History and Philosophy of Science Part A* 52:79–87.

Universality Reduced

Alexander Franklin*†

October 2018

Forthcoming in *Philosophy of Science: Proceedings of the PSA 2018*

Abstract

The universality of critical phenomena is best explained by appeal to the Renormalisation Group (RG). Batterman and Morrison, among others, have claimed that this explanation is irreducible. I argue that the RG account is reducible, but that the higher-level explanation ought not to be eliminated. I demonstrate that the key assumption on which the explanation relies – the scale invariance of critical systems – can be explained in lower-level terms; however, we should not replace the RG explanation with a bottom-up account, rather we should acknowledge that the explanation appeals to dependencies which may be traced down to lower levels.

1 Introduction

While universality is best explained with reference to the Renormalisation Group (RG), that explanation is nonetheless reducible. The argument in defence of this claim is of philosophical interest for two reasons: first, the RG explanation of universality has been touted by Batterman (2000, 2017) and

*alexander.a.franklin@kcl.ac.uk

†I am grateful to Eleanor Knox, and to the audience of the IMPS 2018 conference in Salzburg for helpful comments. This work was supported by the London Arts and Humanities Partnership.

Morrison (2012, 2014) as a significant impediment to reduction. Second, universality is a paradigm instance of multiple realisability (MR) in the philosophy of physics; as such it is regarded as irreducible by those who accept the multiple realisability argument against reduction. My account charts a middle course: I deny claims that RG explanations are irreducible, and I deny that universality is *best* explained from the bottom up.

The view of reduction advocated here is non-eliminativist; the best explanations are often higher-level explanations: such explanations are more parsimonious, more robust, and have broader applicability than lower-level explanations. In general, such higher-level explanations ought not to be replaced by lower-level explanations, rather the parts of theories on which such explanations rely may be understood in lower-level terms; reducible explanations satisfy the following two conditions: (a) each higher-level explanatory dependency is explained by or derived from a lower-level dependency, and (b) the abstractions involved in constructing the higher-level explanations are justified from the bottom up.[1]

In §2 I outline the RG explanation of universality. Although my reductive claims may generalise, I focus exclusively on the field-theoretic approach to the RG.[2] I claim that this explanation follows a general formula for explaining multiply realised phenomena. §3 considers the arguments of Batterman and Morrison, and analyses their force against any putative reduction.

In §4 I note that the RG explanation is a higher-level explanation. As it is less contentious that the common features of each universality class are reducible, I simply assume that that's the case in this paper. The nub of the debate rests on the RG: I show that the RG arguments rely on the assumption of scale invariance and the abstractions engendered by that assumption. I argue that the applicability of this assumption may be explained from the bottom up. Thus, I claim, that my reduction satisfies (a) and (b) above.

[1] While I expect the claims in this paper to be compatible with many different accounts of explanation, they are most straightforwardly cashed out on an interventionist approach – see Woodward (2003).

[2] See Franklin (2018) and Mainwood (2006) for arguments that only this approach provides an adequate explanation of universality.

2 The RG Explanation of Universality

'Universality' refers to the phenomenon whereby diverse systems exhibit similar scaling behaviour on the approach to a continuous phase transition. Continuous phase transitions occur at the critical temperature, a point beyond which systems no longer undergo first-order phase transitions.[3] The approach to this phase transition can be very well described by power laws of the form $a_i(t) \propto t^{\alpha}$ where t is proportional to the temperature deviation from the critical temperature and α is the critical exponent – a fixed number which leads to a characteristic curve on temperature-density plots.[4]

Different physical systems can be categorised into universality classes: members of the same class have identical critical behaviour – the same set of critical exponents $\{\alpha, \beta, ...\}$ for several power laws – while their behaviour away from the critical point and microscopic organisation may be radically different. For example, fluids and magnets are in the same universality class despite otherwise having totally different chemical and physical properties.

Each physical system which exhibits critical phenomena may be described at the critical point by the same mathematical object – the Landau-Ginzburg-Wilson (LGW) Hamiltonian. That Hamiltonian will include the features – the symmetry and dimensionality – which sort these systems into their universality classes. The RG argument demonstrates that the LGW Hamiltonian applies to a wide range of systems at the critical point by showing that any additional operators which may be appended to that Hamiltonian will fall away on approach to criticality, where only the central LGW operators will remain. The following steps are essential to the explanation thus on offer:[5]

1. Define the effective Hamiltonian for your system of interest:
   (i) Specify the order parameter with symmetry and dimensionality.
   (ii) Specify the central operators of the LGW Hamiltonian.

[3] Note that not all continuous phase transitions are associated with first-order phase transitions in this way.

[4] E.g. the specific heat (in zero magnetic field) c scales as $c \sim (t^{-\alpha})/\alpha$ as $t \to 0$ where $t = \frac{T - T_c}{T_c}$.

[5] To see a full account of the physics of universality and details of the RG see Binney et al. (1992) and Fisher (1998); the philosophical aspects of such an explanation are discussed in detail in Batterman (2016) and Franklin (2018).

(iii) Specify operators in addition to the terms in the LGW Hamiltonian.

2. Apply the RG transformations to that Hamiltonian.
3. Examine the flow towards fixed points in the critical region and note that some operators are irrelevant to the critical behaviour.
4. Thus divide the set of operators into subsets: 'relevant', 'irrelevant' and 'marginally relevant'.
5. Repeat for other systems of interest.

In order to explain universality we must identify commonalities between the different systems in the same universality class – 1(i) and 1(ii) above – and show that such commonalities are sufficient for the common behaviour – 2-4 above. Although 1(iii) can't, in general, be done explicitly, the explanation only depends on the RG demonstration that all distinguishing features are irrelevant – it's not necessary to say exactly which those distinguishing features are. As discussed below, the infinities which are central to some of the anti-reductionist arguments feature in steps 3 and 4.

Overall the explanation takes the following form: consider a universality class composed of four different physical systems A-D. Each of A-D is described in step 1 by an effective Hamiltonian; effective Hamiltonians are ascribed to systems on the basis of various theoretical and empirical data. The RG explanation of universality, by virtue of steps 2-4, tells us that all the details which distinguish A-D, i.e. their irrelevant operators, are, in fact, irrelevant to the critical phenomena. Thus we have an explanation for how otherwise different systems exhibit the same phenomena at the critical point. This explanation relies, of course, on the RG transformations which allow for the categorisation of certain operators as irrelevant.

Importantly, this explanation takes the form of a general explanation of multiply realised phenomena: such phenomena are explained if commonalities are identified among the realisers and these are shown to be sufficient for the multiply realised phenomena to occur. Note that such explanations may be higher level and nothing written so far establishes their reducibility.

3 Anti-reductionist Arguments

Batterman (2000, 2017) and Morrison (2012, 2014) offer two arguments in defence of the view that the explanation just outlined is irreducible. The more general argument is that universality, *qua* instance of multiple realisability, is irreducible because multiple realisability requires abstracted explanations of a particular form.

However, one goal of this paper is to demonstrate that just such abstracted explanations may be reducible. Insofar as my reduction of the RG explanation goes through, we are thus faced with a dilemma: either some instances of MR are, in principle, reducible, or universality is not a case of MR. While I would opt for the former horn, nothing in the rest of the paper hangs on that choice.

The second anti-reductionist argument is much more specific to the case at hand and involves various demonstrations that the RG explanation requires infinities which are inexplicable from the bottom up. As noted by Palacios (2017), two different limits are invoked in the case of continuous phase transitions – the thermodynamic limit and the limit of scale invariance. There is an extensive literature on the thermodynamic limit as it appears in first order phase transitions; as I see no salient differences between appeal to this limit in the two contexts, I do not discuss this further here – see e.g. Butterfield and Bouatta (2012) for a reductionist account of that limit.[6]

The second limit is discussed by Butterfield and Bouatta (2012), Callender and Menon (2013), Palacios (2017), and Saatsi and Reutlinger (2018), among others, and these papers undermine claims that continuous phase transitions are irreducible. However, they pay insufficient attention to the specific role played by the RG (and by the limit of scale invariance) in establishing the irrelevance of certain details, and it is this role which is crucial to the anti-reductionist arguments.[7]

For Batterman, the RG is required because it allows us to answer the following question:

[6]The reductionist claims made here are conditional on a successful resolution of such issues.

[7]For example, Saatsi and Reutlinger (2018, p. 473) do not consider a counterfactual of the form 'if a physical system S did not exhibit effective scale invariance at criticality, then S would not exhibit the critical phenomena of any universality class' in their list of counterfactuals which the RG account is supposed to underwrite.

> **MR**: How can systems that are heterogeneous at some (typically) micro-scale exhibit the same pattern of behavior at the macro-scale? ...
>
> if one thinks (**MR**) is a legitimate scientific question, one needs to consider different explanatory strategies. The renormalization group and the theory of homogenization are just such strategies. They are inherently multi-scale. They are not bottom-up derivational explanations.
>
> [Batterman (2017, pp. 4, 14-15)]

As further elaborated below, the RG seems to Batterman to preclude "bottom-up derivational explanation" because it requires the following infinitary assumption:

> This [fixed point] is a point in the parameter space which, under τ [the RG transformation], is its own trajectory. That is, it represents a state of a system which is invariant under the renormalization group transformation. Of necessity, such a fixed point has an *infinite correlation length* and so lies on the critical surface S_∞. The singularity/divergence of the correlation length ξ is *necessary*.
>
> [Batterman (2011, p. 1045), original emphasis]

I accept that the RG formalism makes use of infinite limits. The salient question, to borrow Norton's (2012) distinction, is whether such infinities are approximations which allow one to use the more tractable infinitary mathematics to approximate features of the finite systems, or, alternatively, idealisations which describe a distinct infinite system. Claiming that the infinities are idealisations would preclude reduction because the macroscopic system with infinite properties has features which may not be reductively explained.

As Batterman demonstrates, the RG argument rests on the assumption of the infinite correlation length which generates absolute scale invariance. In §4 I claim that the physical systems under consideration are not absolutely scale invariant: in fact, one may abstract from the details of the underlying system insofar as such systems are effectively scale invariant; thus the infinitary assumption is best viewed as an approximation.

While Morrison (2014, p. 1155) likewise focusses on explanations of MR phenomena, she claims that RG explanations are irreducible for a different, but related, reason: the "RG functions not only as a calculational tool but as the source of physical information as well". Morrison (2012) makes a similar argument in relation to symmetry breaking in the physics of superconductors. She argues that, in both cases, top-down constraints play an essential role in the physical descriptions which thus rules out reduction. In the present context, Morrison's views may be understood as taking the RG invocation of scale symmetry to be a necessary physical assumption which cannot be understood from the bottom up. Below I argue that the effective scale invariance on which the RG rests is, in fact, reductively explicable. As such, no top-down organising principles are required and Morrison's claims are deflated.

4 Reducing the RG Explanation

Arguments for the reducibility of the explanation of universality have primarily been targeted at Batterman's claims that infinities are essential to the models used to describe continuous phase transitions. I do not have space to consider these arguments in any detail. Suffice it to say that, in my view, none succeeds in reducing the principal feature of the renormalisation group – the assumption of scale invariance. Thus I focus on that aspect of the RG, and claim that it, too, is reducible.

Furthermore, with the notable exception of Saatsi and Reutlinger (2018), not much attention has been paid to the explanation of universality *per se*. This, of course, makes a difference for MR-based objections to reduction, which raise doubts that a reductionist account could explain why the same phenomenon is exhibited in multiple different systems.

As far as the physics is currently developed, the RG plays an ineliminable role in the explanation of universality: it is the only mathematical framework available to predict the precise extent of observed universality of critical phenomena. If its application were truly mysterious, if we had no idea why it worked, then, infinity or no infinity, this would provide exactly the right kind of failure of explanation on which the anti-reductionist could hang their arguments.

I argue in the following that the applicability of the RG to systems un-

dergoing continuous phase transitions is not mysterious. The RG exploits effective scale invariance to set up equations which tell us how certain properties vary with respect to the variation of other properties. It is a piece of mathematics whose applicability is deeply physical – where the assumptions invoked in applying the RG do not hold, the RG's predictions go wrong.

In order fully to reduce the RG explanation, one also must consider the common features shared by each member of the same universality class, and argue that these, too, are reducible to aspects of the microphysical description. Such arguments have been given by the reductionists mentioned above. The innovation of this paper lies in reducing the RG framework, and the assumptions on which it relies; thus, given space constraints, I do not consider the reduction of the symmetry, dimensionality and representation by common Hamiltonians.

4.1 Reducing the Renormalisation Group

The RG argument rests on the assumption of scale invariance, and this is crucial to the demonstration that a class of operators are irrelevant at criticality. I claim that we can provide a bottom-up explanation of this scale invariance and that, as such, the RG arguments provide a mathematical apparatus for relating scale invariance to the irrelevance of certain details. One can see, heuristically, how scale invariance relates to universality: if the system at criticality is effectively scale invariant then many of that systems' features – those which are scale dependent – will turn out to be irrelevant at criticality, and all that will remain are those shared features such as the symmetry and dimensionality.

To argue that the RG explanation is reducible, I first give a more general characterisation of an RG flow. The calculation of each system's dynamics involves integration over a range of scales and energies. The highest energy (smallest scale) cutoff (denoted Λ) corresponds to the impossibility of fluctuations on a scale smaller than the distance between the particles in the physical system. The RG transformation involves decreasing the cutoff thereby increasing the minimum scale of fluctuations considered. Iterating this transformation generates a flow through parameter space designed to maintain the Hamiltonian form and qualitative properties of the system in question.

The RG transformation $\mathcal{R}$ transforms a set of (coupling) parameters $\{K\}$ to another set $\{K'\}$ such that $\mathcal{R}\{K\} = \{K'\}$. $\{K^*\}$ is the set of parameters which corresponds to a fixed point, defined such that $\mathcal{R}\{K^*\} = \{K^*\}$. This fixed point corresponds to the critical point defined physically. At the fixed point, the RG transformation (which changes the scale of fluctuations) makes no difference. Thus the fixed point encodes the property of scale invariance.

Given the Hamiltonian of one of our models, one can define an RG transformation which generates a flow that allows one to: (i) classify certain of the coupling parameters of the system in question as (ir)relevant to its behaviour near the fixed point, (ii) extract the critical exponents from the scaling behaviour near the fixed point.

The RG may be understood as a mathematical framework for exploring how certain properties vary with changing energy, length-scale, or, by proxy, temperature, on approach to the scale invariant critical point. Philosophical discussions of the RG are occasionally prone to mysterianism, but the RG should be considered to be no different from, for example, the calculus. As Wilson (1975, p. 674) notes: "the renormalization group ... is the tool that one uses to study the statistical continuum limit [the point of scale invariance] in the same way that the derivative is the basic procedure for studying the ordinary continuum limit".

The Hamiltonian which represents the system at the critical point, from which the critical exponents are extracted, is scale invariant at the fixed point – all the scale dependent contributions have gone to zero. Such Hamiltonians are known as 'renormalisable'. As such, the explanation provided below for the effective scale invariance of physical systems at criticality underlies the fact that such systems are well-described by renormalisable Hamiltonians at fixed points.

My argument has two steps: I demonstrate that scale invariance is implicit in the power law behaviour which is intrinsic to universality; then I provide a bottom-up explanation of the effective scale invariance for liquid-gas systems, a story somewhat motivated by the observation of critical opalescence. Thus, I show how scale invariance features in the mathematics – the Hamiltonian's renormalisability and the power laws, and how it features in the observed physics – the critical opalescence is a direct consequence of the bottom-up story.

The universality of critical phenomena lies in the sharing of power laws,

and hence critical exponents, between members of the same universality class. In what sense are such power laws scale-free? As Binney et al. (1992, p. 20) explain, a phenomenon obeying a power law is independent of scale because one could multiply its characteristic scale length by some factor and the ratio of values will remain constant. For example, consider the power law $f_1 = (r/r_0)^\eta$, and its measurement in the range $(0.5r_0, 2r_0)$. The ratio of largest to smallest value will be identical for measurements centred on $r_0, 10r_0, 100r_0$ – it will always be $4^{|\eta|}$, thus one may superimpose all the power laws by a simple change of scale. By contrast, for $f_2 = \exp(r/r_0)$ the ratio of values will change on scale changes.

Such systems are therefore described as scale-free; the RG is used to predict that at the point of scale invariance the heterogeneous features will be irrelevant. So, in order to work out when this framework is applicable, and why it works, we ought to look at each individual system, (for our purposes let's reserve inquiry to liquid-gas and ferromagnetic-paramagnetic systems) and identify the underlying processes which lead to effective scale invariance at the critical point. The following two caveats apply to this proposal for reduction:

First, it might be objected that universality may only be explained if the same processes are identified across all the systems exhibiting the universal behaviour; if that were so, the strategy employed here would be inadequate. However, universality may be explained by demonstrating that two conditions are fulfilled: that all the systems share common features, and that their heterogeneous details are irrelevant. While it's essential that the common features are shared by all the systems, the mechanism by which the heterogeneities are irrelevant may differ, so long as all the heterogeneities in fact end up as irrelevant.

Second, although the power laws and renormalisable Hamiltonians at the fixed point are absolutely scale invariant, the physical systems will, at best, be effectively scale invariant – that is, scale invariant within a certain range of length-scales. That should be acceptable because we know that scale invariance is never exactly true of a system: any real system will be finite and thus violate the assumption at some scale. Moreover, this will not generate empirical problems because the power laws are observed for systems approaching criticality – they are predictions about $T \to T_c$, not $T = T_c$. Thus one should only assume that critical exponents asymptotically approach those predicted at the fixed point. While infinite assumptions are required in order to impose the full scale invariance for RG analy-

sis, I claim that we can explain effective scale invariance for finite systems, and that absolute scale invariance is an approximation invoked to make the mathematics tractable.

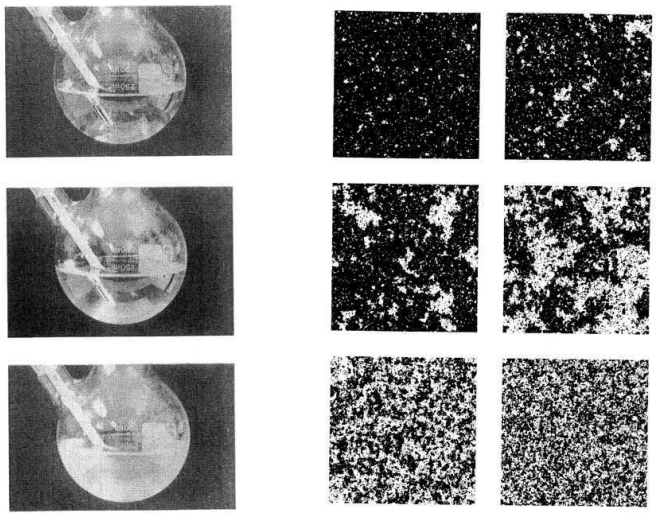
Scale invariance, as it manifests in systems at criticality, is known as 'self-similarity': as scales change the system resembles itself. How do we account for such self-similarity? The critical point, at which a continuous phase transition occurs, corresponds (for liquid-gas systems) to the highest temperature and pressure at which liquid and gas phases can be distinguished.

As is well known, there is a plateau in pressure-volume diagrams, which corresponds to the latent heat (or enthalpy) of vapourisation. This, roughly, is the extra energy needed to break the intermolecular bonds which distinguish liquids from gases and vapours. At the critical point this plateau, and the latent heat of vapourisation vanishes. Now it's difficult precisely to work out the binding energies of the intermolecular bonds. The values for this will be material dependent, and surface tension dependent, and will change at different pressures. But the heuristic argument tells us that the reason the plateau vanishes is because the system has enough temperature, and thus the molecules have sufficient energy to equal the binding energy. The point at which binding energy is exactly matched by kinetic energy will be the critical point.

The isothermal compressibility (κ) is defined as $\kappa_T = \frac{-1}{V}\left(\frac{\partial V}{\partial p}\right)_T$. This corresponds to how much the volume will change (∂V) with a given pressure change (∂p) at fixed temperature (T). As supercritical fluids have far higher compressibility than liquids, and both are present at the critical point, the compressibility diverges. Given, in addition, that the latent heat is zero at criticality, there's nothing to prevent a given bubble expanding arbitrarily. Thus we ought to expect the system to have bubbles of all sizes: this is what is meant by the claim that the system is dominated by fluctuations and has no characteristic scale.[8]

Negligible energy cost for transitions and infinite compressibility leads to self-similarity, and, in certain fluids, the bubbles at all scales lead to a high refraction of visible light. Thus otherwise transparent fluid may become opaque and milky-white. This is known as 'critical opalescence' – see figure 1(a) – and is a visible correlate of a system at criticality.

[8]Note that, for first order phase transitions, the compressibility also diverges; this doesn't lead to scale invariance because latent heat is finite.



(a) (b)

Figure 1: From Binney et al. (1992, pp. 10,19). (a) Critical opalescence is visible when arbitrarily large bubbles form in liquid at criticality. (b) Increasing loss of characteristic scale as $T \to T_c$ in simulations of the Ising model.

Such self-similarity is conceptually crucial to the applicability of the renormalisation group: in order to extract critical exponents from RG equations one identifies a renormalisable Hamiltonian which is scale invariant at the fixed point. Without fluctuations across all scales, systems would fail to be well modelled by such Hamiltonians. The physical argument for diverging fluctuation size justifies the use of a scale invariant mathematical model to represent such systems. Thus, for critical phenomena, the applicability of the RG depends on scale invariance, where this assumption is explicable from the bottom up.

Demonstrating these claims quantitatively is difficult, but the heuristic argument is convincing. Kathmann (2006) reviews theories of the nucleation of gas bubbles in water which generate accurate predictions concerning the rate of bubble growth and the threshold for stability over a range of temperatures; although these models do not reach the critical point, progress is being made.[9]

[9]Constructing exact models is especially difficult because of the fluctuations at a wide range of length scales – precisely the reason that the RG is employed.

Of course, further work could be done to develop these arguments and make them more precise. But there seems to be, in the above, a sound qualitative argument and no in-principle barriers to full derivation. This 'in-principle' ought not to be problematic: we know the relevant physical principles, even if quantitative models are still unavailable.

Moreover, as discussed below, and depicted in figure 1(b), the Ising model allows us quantitatively to predict analogues of the results for liquid-gas systems. While well short of a full explanation, the following discussion illustrates how self-similarity may be reduced for magnetic systems. By treating the Ising model as a stand-in for such systems, a similar kind of reasoning to that given above will go through.

Below the critical point, energy fluctuations will lead to random isolated spin flips. Such flips will be energetically costly and tend to be reversed. The higher the energy, the more likely these are to occur, and if sufficiently many occur then a patch will form, and other spins will have some tendency to align themselves with this patch. However, below the critical point, such patches beyond a certain size will be too costly and spins will overall remain aligned (there is some small probability of net magnetisation flipping, but this is increasingly unlikely further below the critical point).

At the critical point, the energy of the atoms in the lattice is greater than the energetic cost of violating spin alignment, and patches can become arbitrarily large. This results from the latent heat's vanishing and the divergence of the magnetic susceptibility (χ) on approach to the critical point. $\chi_T = \left(\frac{\partial m}{\partial B}\right)_T$ where m is the magnetisation and B represents an external magnetic field. Universality is manifested by the fact that the susceptibility and the compressibility both diverge according to identical power laws with the same critical exponent γ: χ_T, $\kappa_T \sim (T - T_c)^{-\gamma}$. Thus, we have self-similarity and effective scale invariance with bubbles or patches arbitrarily large up to the size of the system.

My aim is to establish the reducibility of the RG relevance and irrelevance arguments. I have demonstrated that the RG is a mathematical procedure that extracts information based on the empirically and theoretically justified assumption of effective scale invariance; this has been shown to be a property shared by different systems at criticality. The key ingredients for effective scale invariance are features of the interactions of neighbouring sub-systems, and the particulate constitution of the materials. While that suggests that these materials are not so different after all, it's worth empha-

sising that the systems which exhibit universal behaviour are nonetheless dissimilar away from the critical point – it's clear that magnets and liquids have many distinct chemical and physical properties.

The assumption of scale invariance plays a crucial role for the RG – it licences the discarding of scale dependent details; it is precisely this discarding of details which ensures that all systems are commonly described at the critical point. Moreover, discarding such details is what gives the higher-level explanation its stability and parsimony. It is thus incumbent on the reductionist to explain how the higher-level RG account is successful despite its leaving out such details. So, the reductionist should identify physical processes at the lower level which ensure the irrelevance of the discarded details.

As argued above, the physical processes in question are exactly those which lead to effective scale invariance. The fluctuations at all scales make it such that the scale-dependent properties which distinguish systems away from criticality are irrelevant at criticality, when the system is effectively scale invariant. We have identified, at the molecular level, the physical mechanisms which prevent variations in the discarded details from leading to changes in the higher-level description of the system. As such, we are assured that the explanatory value of the higher-level explanation is a consequence of features of the lower-level system.

One upshot of this reductionist account is that we may specify the conditions under which the higher-level description remains a good one. The discarded details are irrelevant while the large scale fluctuations – the bubbles or patches – dominate the physics. As we move to systems which are less scale invariant, as the bubbles die down, the critical point becomes a less accurate description and each system in the class will start to exhibit distinct behaviour. This is reflected in the fact that the macroscale RG description only derives the shared behaviour at the fixed point of scale invariance and predicts distinct behaviour away from the fixed point.

I end this section with the following intuitive physical gloss on the RG explanation: "[b]ecause the fluctuations extend over regions containing very many particles, the details of the particle interactions are irrelevant, and a great deal of similarity is found in the critical behavior of diverse systems" (A. L. Sengers, Hocken, and J. V. Sengers (1977, p.42)). Since we can explain the wide-ranging fluctuations from the bottom-up, the RG explanation of universality is reducible.

5 Conclusion

The field-theoretic RG framework, together with the common features of physical systems in the same universality class, explains how those systems all display the same critical phenomena when undergoing continuous phase transitions. That explanation is a higher-level explanation.

That higher-level RG explanation is nonetheless reducible. That is, we may explain in terms of the microstructure of each system how it is that each aspect of the higher-level explanation is explanatory. We may, in particular, show why the RG categorisation of operators as relevant and irrelevant works. That division depends on the assumption of scale invariance, and the assumption of scale invariance is justifiable when systems are effectively scale invariant at criticality.

The anti-reductionist claim that universality is MR, and MR is essentially irreducible has been undermined by demonstrating that we may arrive at a bottom-up understanding of the common features and of what makes such features sufficient for the common behaviour.

The further argument that the use of the infinite limit imposes an irreducible divide between the higher-level and lower-level models has similarly been countered: while we move to the infinite limit in order to make the mathematics simpler, the effective scale invariance can be shown to follow from details of the particle interactions at criticality – that's what identifies the critical point and allows us to make the corresponding abstractions from scale dependent details. Provided with this bottom-up explanation, there is no further reason to claim that the infinite limit is an idealisation rather than an approximation: for we have explained from the bottom up how the system is approximately self-similar.

One upshot of this discussion is that the RG is not to be regarded as mysterious, or, somehow, as the source of physical information. It is applicable only insofar as the systems to which it is applied have the relevant properties, and their having such properties may be reductively explained.

References

Batterman, Robert W. (2000). "Multiple Realizability and Universality". In: *The British Journal for the Philosophy of Science* 51.1, pp. 115–145.

Batterman, Robert W. (2011). "Emergence, singularities, and symmetry breaking". In: *Foundations of Physics* 41, pp. 1031–1050. DOI: 10.1007/s10701-010-9493-4.

— (2016). "Philosophical Implications of Kadanoff's work on the Renormalization Group". In: *Journal of Statistical Physics (Forthcoming)*.

— (2017). "Autonomy of Theories: An Explanatory Problem". In: *Noûs*. DOI: 10.1111/nous.12191.

Binney, James J. et al. (1992). *The Theory of Critical Phenomena: an Introduction to the Renormalization Group*. Clarendon Press, Oxford.

Butterfield, Jeremy and Nazim Bouatta (2012). "Emergence and Reduction Combined in Phase Transitions". In: *AIP Conference Proceedings* 1446, pp. 383–403. DOI: 10.1063/1.4728007.

Callender, Craig and Tarun Menon (2013). "Turn and Face the Strange ... Ch-ch-changes Philosophical Questions Raised by Phase Transitions". In: *The Oxford Handbook of Philosophy of Physics*. Ed. by Robert W. Batterman. Oxford University Press, pp. 189–223.

Fisher, Michael E. (1998). "Renormalization group theory: Its basis and formulation in statistical physics". In: *Reviews of Modern Physics* 70.2, p. 653.

Franklin, Alexander (2018). "On the Renormalization Group Explanation of Universality". In: *Philosophy of Science* 85.2. DOI: 10.1086/696812.

Kathmann, Shawn M. (2006). "Understanding the chemical physics of nucleation". In: *Theoretical Chemistry Accounts* 116.1, pp. 169–182. DOI: 10.1007/s00214-005-0018-8.

Mainwood, Paul (2006). "Is More Different? Emergent Properties in Physics". PhD thesis. University of Oxford.

Morrison, Margaret (2012). "Emergent Physics and Micro-Ontology". In: *Philosophy of Science* 79.1, pp. 141–166. DOI: 10.1086/663240.

— (2014). "Complex Systems and Renormalization Group Explanations". In: *Philosophy of Science* 81.5, pp. 1144–1156. DOI: 10.1086/677904.

Norton, John D. (2012). "Approximation and Idealization: Why the Difference Matters". In: *Philosophy of Science* 79.2, pp. 207–232.

Palacios, Patricia (2017). *Phase Transitions: A Challenge for Reductionism?* URL: philsci-archive.pitt.edu/13522/.

Saatsi, Juha and Alexander Reutlinger (2018). "Taking Reductionism to the Limit: How to Rebut the Antireductionist Argument from Infinite Limits". In: *Philosophy of Science* 85.3, pp. 455–482. DOI: 10.1086/697735.

Sengers, Anneke Levelt, Robert Hocken, and Jan V. Sengers (1977). "Critical-point universality and fluids". In: *Physics Today* 30.12, pp. 42–51.

Sober, Elliott (1999). "The multiple realizability argument against reductionism". In: *Philosophy of Science*, pp. 542–564.

Wilson, Kenneth G. (1975). "The renormalization group: Critical phenomena and the Kondo problem". In: *Reviews of Modern Physics* 47.4, pp. 773–840.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. Oxford University Press.

Title: There Are No Ahistorical Theories of Function

Author: Justin Garson

Abstract: Theories of function are conventionally divided up into historical and ahistorical ones. Proponents of ahistorical theories often cite the *ahistoricity* of their accounts as a major virtue. Here, I argue that none of the mainstream "ahistorical" accounts are actually ahistorical. All of them embed, implicitly or explicitly, an appeal to history. In Boorse's goal-contribution account, history is latent in the idea of statistical-typicality. In the propensity theory, history is implicit in the idea of a species' natural habitat. In the causal role theory, history is required for making sense of dysfunction. I elaborate some consequences for the functions debate.

Keywords: Philosophy of biology; biological function; selected effects; causal role; fitness contribution

Address: Department of Philosophy, Hunter College of the City University of New York, 695 Park Ave., New York, NY 10065

Email: jgarson@hunter.cuny.edu

1. Introduction

Theories of function are conventionally divided up into two main categories, historical and ahistorical (or backwards-looking and forwards-looking). The selected effects theory (Neander 1983, 1991; Millikan 1984) is an example of a *historical* theory, but there are other historical theories, including some versions of the organizational theory (McLaughlin 2001), and the weak etiological theory (Buller 1998). *Ahistorical* theories include Boorse's goal-contribution account (1976; 1977; 2002), the propensity theory (Bigelow and Pargetter 1987), and the causal role theory (Cummins 1975; Hardcastle 2002; Craver 2001; 2013). In the 1970s and 1980s, it was common to see these two sorts of theories as competing with each other, though more recently, philosophers of biology have generally adopted a pluralistic stance, and see them as capturing different aspects of real biological usage (OMITTED). Still, the validity of the basic distinction has never been seriously challenged.

Many proponents of ahistorical theories have argued that we should accept their theories precisely *on account of* their being ahistorical. In other words, their alleged ahistoriticity is often held up as a significant virtue of their theories, and a strong reason to prefer them to historical theories (or at least a strong reason to think they capture a significant strand of ordinary biological usage). There are two arguments along these lines. The first argument appeals to bald intuition, and says that it's just obvious that functions don't always need history. One fanciful variant of this argument appeals to science fiction cases, like swamp creatures, instant lions, and randomly-generated worlds (e.g, Boorse 1976, 74; Bigelow and Pargetter 1987, 188). But one doesn't have to go as far as science fiction to find plausible cases of ahistorical functions in biology. Many philosophers have a strong intuition that, the very first time a new biological trait emerges and begins to benefit the organism, it has a *function* even if it was never selected for (e.g., Boorse 2002, 66; Bigelow and Pargetter 1987, 195; Walsh and Ariew 1996, 498). The second argument, which is closely related, appeals to ordinary biological usage, not intuition. It says that historical theories run against the way biologists ordinarily think and talk about functions. At least sometimes, when biologists attribute functions to traits, they do not *cite* or *refer to* or *think about* history or evolution (e.g., Godfrey-Smith 1993, 200; Amundson and Lauder 1994, 451; Walsh 1996, 558; Boorse 2002, 73). Hence, ahistorical theories capture important strands of real biology.

In light of the above, my thesis might come as a bit of a shock. I claim that *there are no ahistorical theories of function* – or, to put it more precisely, the mainstream versions of the allegedly ahistorical theories on the market are not actually ahistorical. If we poke and prod at those theories a bit, a historical element falls out, like contraband stashed away in a suitcase. In Boorse's version of the goal-contribution account, history is explicitly embedded in his notion of a *statistically-typical* contribution to fitness. In the propensity account, history is embedded, a little less explicitly, in the idea of a species' *natural habitat*. Finally, I claim that the only way the causal-role theorist can hope to make sense of dysfunction is to appeal to history.

If this thesis is correct – that there are no ahistorical theories of function – three consequences immediately follow. First, we need to jettison this whole way of dividing up theories of function. The distinction between etiological and non-etiological theories serves us much better, as I'll describe in the conclusion. The distinction between etiological and non-etiological theories doesn't map onto the distinction between historical and ahistorical theories; rather, *these are two ways of being historical*. Second, given that there are no ahistorical views, a good portion of the arguments that have been put forward to date for these theories (those I mentioned above) are unsound. A third consequence is that one popular way of thinking about function pluralism must fail. This sort of pluralist wishes to sort all biological usage under two main umbrella theories, the selected effects theory and the causal role theory. An argument for this sort of pluralism is that it mirrors the two main uses of "function" in biology, the historical sense and the ahistorical sense. If I'm right, this incarnation of the pluralist project can't possibly work.

Before I move on, there is one big qualification I must get out of the way. One could, just for fun, *invent* a purely ahistorical theory of function. One could assert, for example, that *all* of a trait's effects are its functions. This theory (pan-functionalism?) would be ahistorical, to be sure, since even if the world were created two seconds ago in pretty much its present form, things would still have effects, and so they'd still have functions. In fact, sometimes scientists actually *do* use the word "function" synonymously with "effect." They say things like, "climate change is a *function* of deforestation," or "poor academic performance is a *function* of malnutrition." Clearly, there are some ahistorical uses of "function." But this isn't the ordinary biological use, which the theories I cite above are trying to capture.

So, I need to amend my thesis slightly. Instead of saying that there are no ahistorical theories of function, I want to say that any theory of function that satisfies two very minimal, very traditional, and largely uncontroversial, adequacy conditions, is *also* a historical theory. First, the theory should capture some distinction between functions and accidents (the function of the nose is to help us breathe but not hold up glasses). Second, the theory should capture the possibility of malfunctioning or dysfunction. If my heart seizes up due to cardiac arrest, it's failing to perform its function or it's dysfunctional. All of the theorists I engage with in this paper purport to satisfy these two adequacy criteria, or something like them, so I'm not begging any questions by insisting on these conditions.

Here's the plan for the rest of the paper. There are five sections. After the introduction, I'll turn to Boorse's version of the goal-contribution theory, and show how it explicitly contains a historical element (Section 2). Then I'll turn to the propensity theory and show how it contains a reference to history, buried inside the idea of a trait's *natural habitat* (Section 3). I will then show how the causal-role theory, if it is to make any sense of dysfunction, must include a reference to history (Section 4). In the conclusion (Section 5), I'll reiterate the big consequences for thinking about functions and suggest a better way of dividing up theories of function.

2. Boorse's Goal-Contribution Account

Boorse's view (1976; 1977; 2002), at the most general level, is a goal-contribution account. It holds that a trait's function is just its contribution to a goal. The plausibility of this view stems from its ability to reconcile artifact and biological functions in a single theory: the function of an artifact depends on its contribution to the goal of its user; the function of a biological trait depends on its contribution to the goal of the organism or the lineage. Here, I'll focus on the subclass of functions he calls *physiological* functions.

For Boorse, the *physiological* function of a trait is its species-typical contribution to the survival and reproductive prospects of an organism (1977, 555; 2002, 72). (To be more precise, Boorse carves up species into subgroups based on age and sex; the function of a trait is its typical contribution to fitness within the members of that subgroup.) Though he doesn't define a corresponding notion of *dysfunction*, he defines a closely related notion of *disease*: a disease is simply a state that "reduces one or more functional abilities below typical efficacy."

One of Boorse's arguments for the superiority of his theory over Wright's (1973) etiological approach, and the selected effects theory of Millikan (1984) and Neander (1983), is that his approach *makes no reference to history*. He advances two arguments for the value of this ahistorical approach; one appeals to ordinary biological usage, and the other appeals to intuition. First, he says, the goal-contribution account fits ordinary biological usage: "in talking of physiological functions, they [that is, pre-Darwinian biologists] did not mean to be making historical claims at all. They were simply describing the organization of a species as they found it" (1976, 74). The same is true of current physiologists, who have "*no thought* of explaining [a trait's] history" when they assign functions to them (Boorse 2002, 73, emphasis mine). All historical theories of function simply miss how physiologists have always used the word "function." His second argument appeals to intuition. He says that intuition revolts against putting history into functions, as attested to by his instant lions case. If the lion species sprang into existence by "unparalleled saltation," one would *not* say that the parts of lions don't have functions (ibid.; also see Boorse 2002, 75). Again, functions can't be historical.

Neander (1991, 182) raised a now-famous objection against Boorse; she pointed out that Boorse's view, as it stands, can't make sense of pandemic disease: "dysfunction can become widespread within a population…A statistical definition of biological norms implies that when a trait standardly fails to perform its function, its function ceases to be its function; so that if enough of us are stricken with disease (roughly, are dysfunctional) we cease to be diseased, which is nonsense." Pandemic diseases, moreover, don't just occupy the realm of science fiction, as in P. D. James' *The Children of Men*. UV radiation poisoning in anurans is a good example of pandemic dysfunction. Sadly, climate change might create many more pandemic dysfunctions very soon. A good theory of function should at least allow for the *conceptual* possibility that all, or most, tokens of a certain trait in a certain species are dysfunctional (or as Boorse prefers, "diseased").

Intriguingly, Boorse doesn't deny the possibility of pandemic disease. Instead, he says that in order to make sense of pandemic disease, one has to appreciate function's

historical depth. Specifically, he says that when we consider what is "statistically typical" for a trait, we cannot just look at what is typical right now. Rather, we have to consider what is typical within a long slice of time that extends far back into the past: "Obviously, some of the species' history must be included in what is species-typical. If the whole earth went dark for two days and most human beings could not see anything, it would be absurd to say that vision ceased to be a normal function of the human eye (2002, 99)." He tells us that this time-slice should be longer than "a lifetime or two," and might include "millennia."

This is an extraordinary admission, given that much of Boorse's core argument *for* his view was propped up on the claim that both biology and intuition need purely ahistorical functions, uncluttered by history. His admission implies that two of his key arguments for the view (cited above), are unsound. First, by his own admission, it's not the case that biologists don't refer to history; implicitly, when they talk about what's statistically-typical, they *are* talking about history. Second, regardless of whether or not intuition supports ahistorical functions, Boorse's theory doesn't. It's just not true, on Boorse's account, that if lions popped into being from an unparalleled saltation, their parts and processes would have functions. They wouldn't, since they don't have the right history (or to be more precise, they have no history at all). True, Boorse's history isn't the same *kind* of history that features in the selected effects theory, since it doesn't refer specifically to etiology, but it's still history, and so his arguments that appeal to the ahistoricity of his theory don't work.

3. The Propensity Theory

Bigelow and Pargetter (1987) also developed an influential "ahistorical" theory of function, the propensity theory. They reject the selected effects theory (and etiological accounts more generally) because the selected effects theory gets the *modality* of functions wrong. In other words, the statement, "functions are selected effects," if true, is contingently true; it might be true on the actual world, but there are possible worlds at which it's false. To illustrate the point, they ask us to consider a world that is pretty much the same as ours except that it randomly popped into being five minutes ago. On that world, they claim, there would still be functions, just no selected effects (188): "we have the intuition that the concept of biological function…[is] not thus contingent upon the acceptance of the theory of evolution by natural selection." This consideration prompts the need for an ahistorical theory.

For Bigelow and Pargetter, functions are propensities, or probabilistic dispositions. We might quibble over what exactly dispositions are, but any good definition will cite three parts: structure, environment, and behavior. Consider the solubility of salt. There is a *structure*, namely, the polar molecular structure composed of sodium and chloride; there is an *environment*, namely, water; there is a *behavior*, namely, dissolving. When we say that salt is disposed to dissolve in water, we're saying that, if you were to take this structure, and put it in this environment, it would perform this behavior.

Functions, too, are dispositions. Consider "the function of the heart is to circulate blood." For this statement to be true, there must be a structure (the heart, embedded the right way in the circulatory system), an environment (which they call the creature's *natural habitat*), and a behavior (conferring a fitness boost on the organism). If one were to put the structure in its natural habitat, it would increase the fitness of the organism (relative, I suppose, to creatures without hearts). The crucial distinction between their view and Boorse's is that in their view, a trait's function doesn't depend on actual frequencies of performance. A trait needn't have an actual track record of boosting fitness to have a function; a mere propensity will do.

This raises the thorny question of what a creature's *natural habitat* is. For they're clear that a creature's natural habitat isn't just any environment the creature happens to find itself in. Unfortunately, they refuse to define this crucial notion; instead, they brush it off as vague, but unproblematically so: "there may be room for disagreement about what counts as a creature's 'natural habitat;' but this sort of variable parameter is a common feature of many useful scientific concepts" (192). But one could at least form the suspicion that if one analyzed this unproblematically vague notion, one would find some reference to history tucked away inside of it.

This suspicion is confirmed in the very next paragraph. There, they tell us that, if a creature's environment were to change very suddenly, then "natural habitat" will still refer to the *old* environment, and not the *new* one (ibid). There's a time lag built into the very idea of a natural habitat. So, for example, if climate change melts enough Arctic ice, then, at least for a time, the polar bear's natural habitat (and by extension, the natural habitat of the trait itself, namely, their thick, water-repellant fur) is the icy habitat of yore and not the contemporary, denuded one. They take that as given, and I agree.

But why would this be? What *makes it the case* that this is true, namely, that in cases of rapid habitat change, "natural habitat," at least for a time, refers to the old environment and not the new one? What makes it true, I suspect, is that the idea of a natural habitat is an intrinsically historical notion. It's something like the environment within which the organism recently survived and thrived. And if that's not what a natural habitat is, I would like to know what it is *such that*, if a creature's actual habitat shifts suddenly, the natural habitat is still the old one. Just because a concept is vague around the edges, that doesn't exempt one from the obligation to give some sort of analysis.

Hence, I conclude that, contrary to rumor, the propensity theory is not an ahistorical theory, or not demonstrably so. But if that's right, they lose one of the main virtues of the view, which is to get the modality of functions right. To be fair, there's still a sense in which their view *is* ahistorical. What they can do, that the selected effects theorist can't, is to attribute functions to novel traits – so long as that novel trait belongs to the members of a species that has been around long enough to have a natural habitat. Suppose a gene mutation confers a benefit on an organism, say, pesticide resistance on a flour beetle. I suppose they can say that, at the very moment at which it first confers that benefit, the gene mutation has a function, namely, to make the beetle withstand a certain pesticide. This result, they claim, is "intuitively comfortable" (195). But they can say that only

because flour beetles themselves have a history, and so we can talk meaningfully about their natural habitats. Moreover, I think they'll still have a very hard time dealing with dysfunction (Neander 1991, 183), as I hope to show in the next section. Finally, I think there are good theory-neutral reasons for saying that beneficial traits, on their very first appearance, don't have functions, but rather, whatever benefit they bring is an accident. But I won't argue for that here (see OMITTED).

4. The Causal Role Theory

What about the causal role theory of function? This appears to be a purely ahistorical view. The causal role theory says, roughly, that the function of a *component* of a system consists in its contribution, in tandem with the other components, to a system-level capacity of interest (Cummins 1975; Craver 2001; Hardcastle 2002). Craver (2001; 2013) helpfully elaborates this view by specifying that the part in question must be a component of a *mechanism*. All of the basic ingredients of this theory are ahistorical: capacities, components, organization, hierarchy, interests. Even if the world were created five minutes ago, in pretty much its present form, things would still have causal role functions.

The problem enters when we think about dysfunction. Cummins (1975, 758) insisted that functions are dispositions, or capacities: "…to attribute a function to something is, in part, to attribute a disposition to it." The function of a trait *token*, then, consists in its capacity to contribute to a system-level effect. But what if the token in question, through defect or disease, loses the capacity, and so can't contribute to the system-level effect? Then, by Cummins' analysis, it doesn't have the relevant function – so it can't dysfunction either.

Causal role theorists have, by and large, been silent about how to make sense of dysfunctions from this perspective. Almost everything they've had to say on that score, however, is consistent with the following theme: a trait *token* dysfunctions when it can't do what other trait tokens generally, or typically, do to contribute to the system-level effect of interest. Consider Godfrey-Smith (1993, 200): "Although it is not always appreciated, the distinction between function and *mal*function can be made within Cummins' framework…If a token of a component of a system is not able to do whatever it is that other tokens do, that plays a distinguished role in the explanation of the capacities of the broader system, then that token component is malfunctional." Craver (2001, 72), offers the same general line: "…the ascription of a function to a malformed or broken part is derivative upon a description of how that *type* of part (X) fits into a *type* of higher-level mechanism (S). The malformed and broken part can be identified as an X by the typical properties and activities of Xs...." This is, at root, to rely on a statistical norm for making sense of dysfunction.

This account of dysfunction, like Boorse's, stumbles when it encounters the problem of pandemic dysfunction (Neander 1991). For the modification suggested above implies that, if everyone's heart seized up at once, nobody's heart would have a function anymore, so nobody's heart would be dysfunctional. The best way to solve this problem,

and perhaps the only way, is the way Boorse took, namely, to say that the function of a trait is its typical contribution to some system effect, when what's typical is assessed over a chunk of time that stretches back into the past, for at least "a lifetime or two," and perhaps "millennia." But if causal role theorists take that line, they'd have a historical theory.

Craver (2001) and Hardcastle (2002) suggest, all too fleetingly, a different way of thinking about dysfunction, one that depends not on statistics, but on our values and goals, that is, the values and goals of people who make function attributions. Craver (2001, 72) suggests that traits dysfunction when they cannot do what people *want* them to do: "the mechanistic role of the broken part only appears against the fixed backdrop of shared assumptions about a type of mechanism within which parts of this type generally (or preferably) make important contributions." The parenthetical remark alludes to a substantially new doctrine, one that demands our full concentration. It suggests that dysfunction is a mirror of human preferences and goals, of our wishing and wanting. If my heart seizes up, it's dysfunctional, since it's not doing *what I want it to do.*

Hardcastle (2002) makes remarks along similar lines. She first says that the function of a trait - what it's "supposed to do," as she puts it - depends on the goals of the scientific discipline that makes the investigation: "The teleological goal for some trait…depends upon the discipline generating the inquiry" (153). The palmomental reflex causes a chin twitch when you stroke an infant's palm; it's just an accident of cortical wiring with no deep evolutionary rationale. Still, she says, it has the *function* of indicating the state of brain development in infants, because that's how biomedical researches use it. She then says that something malfunctions just when it cannot do what it's supposed to do (152). The palmomental reflex malfunctions when it can't indicate the state of brain development. Simply put, dysfunction happens when a trait can't do what we want.

But dysfunctions cannot be reduced to preferences in any straightforward way; this is a point that's been taken for decades (e.g., Boorse 1977, 544; Wakefield 1992, 372), for reasons that scarcely need to be rehearsed. I'd prefer not to need sleep and water; I'd prefer if nobody had to go through the pain of childbirth or teething, either. But none of those things are diseases or dysfunctions. For that matter, I'd prefer if my hands were equipped with retractable adamantium claws. The fact that my hands can't do what I want them to do doesn't make them dysfunctional. If one really wanted to run with this value-centered line about dysfunction, one would *at least* have to add that, in order for a trait to dysfunction, it's not enough that it doesn't do what I prefer, but I must also have a *reasonable expectation* that it *should* act in the way that I prefer. But what could possibly ground a *reasonable expectation* that my hand (say) work in a certain way? Only this: that hands usually *do* work in the preferred way. But then we're back to statistical norms, and long historical slices of time. This value analysis of dysfunction isn't a contender to a statistical analysis; instead, the former presupposes the latter.

I've walked through three allegedly ahistorical theories of function, and shown that none of them are purely ahistorical; they're tainted with history. The conclusion will say what we should do next.

5. Conclusion

There are no ahistorical theories of function, at least among those that are usually put forward as ahistorical. The first, Boorse's goal-contribution theory, explicitly refers to what is statistically typical for a trait, where what's typical is assessed over a long historical period of time. The second, the propensity theory, refers to the creature's natural habitat, which is implicitly historical. And the third, the causal role theory, can't hope to make sense of dysfunction (or so I argue) without appealing to a statistical norm, and thereby (following Boorse) to history. *No* theory of function will give functions to the parts of swamp creatures, instant lions, or anything on worlds that are similar to ours except for being randomly generated five minutes ago. The propensity theory, at least, can give functions to novel traits as soon as those traits begin benefiting their bearers, as long as the population in which the traits emerge has been around for long enough to have something like a natural habitat. But even that theory will probably encounter problems when it comes to making sense of dysfunction, though I haven't pushed that line in any detail here.

Three immediate consequences follow from this fact. The first is that we should stop dividing up theories of function in terms of historical and ahistorical. The second is that many of the main arguments for the allegedly ahistorical theories are unsound. Third, one popular form of pluralism, which says that there are two main theories of function, corresponding to historical and ahistorical uses of "function" in biology, is untenable.

But if we can't rely on the historical/ahistorical distinction as a way of dividing up functions, how should we talk about them? I think it's best to divide them up into etiological and non-etiological (as theorists are sometimes wont to do anyway). But there's a crucial clarification in order: to say a theory is etiological isn't *just* to say it's historical. It's to say that the theory deals specifically with causal history. The theory purports to capture the sense in which, when we attribute a function to a trait, we're trying to give a causal explanation for why the trait exists. Most other theories of function are non-etiological, in that they do not purport to explain, in a causal sense of "explain," why the trait exists. But they're still historical.

There's a twist to this story. I think there *are* ahistorical theories of function. Consider that climate change is a function of deforestation, poor academic performance is a function of malnutrition, and wildlife habitat is a function of soil. These notions are *ahistorical* through and through. "Function," in this context, means little more than "effect," and perhaps (as in the last of the three examples) "helpful effect." But this tepid sense of function isn't going to sustain a distinction between function and accident, nor will it give us any sense of dysfunction. This is the sort of "function" that Bock and von Wahlert (1965, 274) were getting at when they equated functions with "all physical and chemical properties arising from [the trait's] form." It's also the sort of "function" that Neander (2017) describes in her recent discussion of "minimal functions." But the proponents of the allegedly ahistorical theories want functions to do much more than that. They are trying to capture the ordinary biological sense (or *an* ordinary biological sense)

of "function," where functions differ from accidents and sometimes things dysfunction. Unfortunately, they can't have what they want.

References

Amundson, R., and G. V. Lauder. 1994. Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy* 9: 443-469.

Bigelow, J., and Pargetter, R. 1987. Functions. *Journal of Philosophy* 84: 181-196.

Bock, W. J., and von Wahlert, G. 1965. Adaptation and the form-function complex. *Evolution* 19: 269-299.

Boorse, C. 1976. Wright on functions. *Philosophical Review* 85: 70-86.

Boorse, C. 1977. Health as a theoretical concept. *Philosophy of Science* 44: 542- 573.

Boorse, C. 2002. A rebuttal on functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M. Perlman, 63-112. Oxford: Oxford University Press.

Buller, D. J. 1998. Etiological theories of function: A geographical survey. *Biology and Philosophy* 13: 505-527.

Craver, C. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Craver, C. 2013. Functions and mechanisms: A perspectivalist view. In *Function: Selection and Mechanisms*, ed. P. Huneman, 133-158. Dordrecht: Springer.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72: 741–765.

Godfrey-Smith, P. 1993. Functions: Consensus without unity. *Pacific Philosophical Quarterly* 74: 196-208.

Hardcastle, V.G. 2002. On the normativity of functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M Perlman, 144-156. Oxford: Oxford University Press.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Neander, K. 1983. *Abnormal Psychobiology*. Dissertation, La Trobe.

Neander, K. 1991. Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.

Neander, K. 2017. Functional analysis and the species design. *Synthese* 194: 1147-1168.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47: 373–388.

Walsh, D.M. 1996. Fitness and function. *British Journal for the Philosophy of Science* 47: 553-574.

Walsh, D. M., and A. Ariew. 1996. A taxonomy of functions. *Canadian Journal of Philosophy* 26: 493-514.

Wright, L. 1973. Functions. *Philosophical Review* 82: 139-168.

What do molecular biologists mean when they say 'structure determines function'?

Gregor P. Greslehner*

University of Salzburg & ERC IDEM, ImmunoConcept, CNRS/University of Bordeaux

October 2018

Abstract

'Structure' and 'function' are both ambiguous terms. Discriminating different meanings of these terms sheds light on research and explanatory practice in molecular biology, as well as clarifying central theoretical concepts in the life sciences like the sequence–structure–function relationship and its corresponding scientific "dogmas".

The overall project is to answer three questions, primarily with respect to proteins: (1) What is structure? (2) What is function? (3) What is the relation between structure and function?

The results of addressing these questions lead to an answer to the title question, what the statement 'structure determines function' means.

*Email: gregor.greslehner@gmail.com

Keywords: philosophy of biology, molecular biology, protein structure, biological function, scientific practice

1 Introduction

'Structure' and 'function' are abundantly used terms in biological findings. Frequently, the conjunct phrase 'structure *and* function' or the directional phrase '*from* structure *to* function' is to be found, indicating that there is a special relation connecting these two concepts. The strongest form of this relation is found in the frequent statement that 'structure *determines* function'. One could easily list several hundreds of references containing such phrases. However, in order not to blow up the references section, I will refrain from doing so. Suffice it to say that biologists make highly prominent use of these concepts in describing their research—molecular biologists, in particular. In this paper, I attempt to clarify these concepts, address their relation, and discuss the role they play in molecular biology's explanatory practice. While these issues can be addressed for many different biological entities on different levels of organization, I restrict the discussion primarily to proteins.

What do biologists refer to when they use this phrase? Is there a particular scientific program or strategy behind the slogan 'structure determines function'? Despite the frequent use of this phrase and the concepts to which it refers, a rigorous analysis is missing. Thus, a philosophical clarification would be a valuable contribution to the conceptual foundations of biology. One such fundamental concept is the sequence–structure–function relationship. "The relationships between sequence, structure, biochemical function and biological role are extremely ill-defined and scant

high quality data are available to allow us to analyse them." (Sadowski and Jones, 2009, 360)

In this paper, I attempt to close this gap by developing an explication of both concepts of *structure* and *function* as they are used in biological practice and discussing which relation holds between them. The third component in this "trinity of molecular biology"—sequence—is the least in need of explication. The standard textbook view holds that sequence determines structure, and structure determines function. I will focus on the second relation.

Without reviewing the rich history of these concepts throughout biology at this point, it is worth noting that functionality and form or structure were thought to be intimately linked from early on. In the early days of biology at the macroscale, the structures had to be observed with the naked eye. Thus, the first examples about the form of bodies or their parts and their functions can be found in physiology and anatomy, for example Harvey's notion of the heart's function to pump blood. From the scale of physiology to the molecular scale, structure and function are closely related. What exactly links these two concepts? Is it a determination relation? And if so, which one is determining the other?

With the invention of microscopes and later the emergence of molecular biology, the structures and functions under consideration shifted from macroscopic entities to individual molecules. In fact, molecular biology put the three-dimensional shape of molecules center stage for explaining biological phenomena. This is the focus of this paper. In particular, the discussion will be confined to the structure and function of *proteins*—with special emphasis on the question whether the former determines the latter.

2 The ambiguity of 'structure'

In a first approximation, 'structure' and 'function' could be interpreted as the most general or neutral way of describing what molecular biologists are doing in their research and what their findings are about. These include mainly the three-dimensional shapes of molecules (or larger cellular structures) and the activities (functions) theses molecules perform in living cells, biochemical pathways, chemical reactions, or just individual steps in such mechanisms. The ultimate aim is to explain biological phenomena with molecular mechanisms, whose entities can be described in physical and chemical terms. The structure of molecules can be described in terms of physics and chemistry—function, however, is a concept that does not appear in physics or chemistry. Let's start by taking a closer look at the notion of structure.

'Structure' is an ambiguous term. Applied to proteins, there is the usual nomenclature of *primary structure* (i.e., a protein's amino acid sequence), *secondary structure* (i.e., common structural motifs like α-helices and β-sheets), *tertiary structure* (i.e., the three-dimensional shape of a single folded amino acid chain), and *quaternary structure* (i.e., the final assembly of a protein if it consists of more than one amino acid chain). Other structurally important components are post-translational modifications and prosthetic groups which are not part of its amino acid composition. All these notions of structure have in common that they are about the molecular composition and shape of a molecule. One meaning of 'structure' denotes the sequence of a polymer, the other meaning is about the three-dimensional shape of a molecule. As will be discussed below, another important ambiguity of 'structure' allows to denote the organization of an interaction network. That leaves us with three different meanings of 'structure':

(1) the sequence of a polymer, (2) the three-dimensional shape of a molecule, and (3) the network organization of several biological entities.

While meanings (2) and (3) are candidates for being functional entities, structure as sequence (1) rather relates the sequences of different polymers (DNA, RNA, and proteins) and also plays a central role in determining the three-dimensional shape of a molecule, structure (2). The primary structure of a protein is just the sequence of amino acids that are put together to form a polypeptide. This amino acid sequence is determined by the corresponding protein-coding gene, which is first transcribed into mRNA and then translated into protein by the ribosome. This scheme is known as the "central dogma of molecular biology":

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$$

The arrows might be interpreted as determination relations. The textbook view of protein structure and function proceeds as follows:

$$\text{nucleotide sequence} \rightarrow \text{amino acid sequence} \rightarrow \text{protein structure} \rightarrow \text{protein function}$$

Strong evidence supporting the claim that the three-dimensional shape of a protein is determined by the sequence of amino acids alone was provided by the experiments of Christian Anfinsen, showing that ribonuclease could, after treatment with denaturing conditions, regain its form and function (Anfinsen et al., 1961). Later, Merrifield showed that an *in vitro* synthesized sequence of amino acids can carry out the enzymatic activity of ribonuclease, thus gaining its functional form without the aid of any other cellular component (Gutte and Merrifield, 1971). From this and similar experiments, Anfinsen

built general rules of protein folding as a global energy minimum which depends solely on the sequence of amino acids (Anfinsen, 1973). This view is known as "Anfinsen's dogma".

In 1958, John Kendrew's lab determined the first actual three-dimensional form of a protein, myoglobin (Kendrew et al., 1958). The predominant technique to determine protein structures is still X-ray crystallography (Mitchell and Gronenborn, 2017). Other techniques include nuclear magnetic resonance, cryogenic electron microscopy, and atomic force microscopy. X-ray structures in particular have been supporting the view that there is a unique rigid shape—the protein's native, functional state—which would be necessary and sufficient for a protein to carry out its biological function.

To make a long story short, the relation between nucleotide sequence and amino acid sequence has been generally confirmed (although there are much more complicated mechanisms to it, e.g., splicing). However, the part concerning the protein shape and function proves to be much more problematic. That poses a challenge to what Michel Morange calls "the protein side of the central dogma" (Morange, 2006).

To get from amino acid sequence to three-dimensional structure is known as the *protein folding problem*. As the term 'problem' suggests, it poses a serious challenge and remains unsolved to this day. Even though knowledge-based techniques to predict protein structures from their sequence have become impressively sophisticated, successful, and reliable, there are good reasons to suspect that the protein problem might remain unsolved in principle—if the aim is to predict protein folding based on chemical and physical principles only.

Every two years the best prediction tools are tested in a contest, the Critical Assessment of protein Structure Prediction (CASP). Based on experimentally determined structures which are only published after the participants of the contest have

submitted their predictions, the predictions are then compared to the experimental structure. A similar contest for predicting the functions of proteins exists (Critical Assessment of Functional Annotation, CAFA), although it is much less developed. But what is function in the first place?

3 The ambiguity of 'function'

'Function' is also an ambiguous term (Millikan, 1989)—even more so than 'structure'. There is a rich history of debates surrounding different notions of function. The term 'function' has a long tradition in biology and its philosophy (Allen, 2009). Starting with Aristotle, activities in biology were interpreted to *have a purpose*, to be goal-directed (teleological). The standard example is that the heart's function is to pump blood. That the heart also produces noise is not considered to be functional. Classic accounts of function have been predominantly trying to capture the teleological aspect, for example (Wright, 1973). However, intentionality is a problematic notion in biology. In another important account, Robert Cummins (1975) stressed the importance of a component's contribution to the system in which it is contained, rather than why natural selection has favored a certain trait. Although it makes sense in evolutionary biology to have an account of function that captures the evolutionary developments, molecular biology and protein science operate with a different notion of function, i.e., mainly biochemical activity. There seem to be two entirely different questions: What is a structure doing? And how did this structure evolve to do what it does?

Arno Wouters distinguishes four notions of biological function (Wouters, 2003): (1) (mere) activity, (2) biological role, (3) biological advantage, and (4) selected effect.

The last two are issues of evolutionary biology, whereas the former two fall within the molecular biologist's domain. If function is to be determined by a molecule's three-dimensional shape or organization network, only (1) and (2) seem to be the proper reading of 'function' in this context.

Which entities have functions within living organisms? Depending on the level of organization at which one is operating, one could give a different answer: molecules, organelles, cells, tissues, organisms, individuals, populations, ecosystems. The most prevalent candidates in molecular biology are certainly DNA and proteins, although lipids and other biomolecules play important roles in life processes, too.

Traditionally, functions have been attributed to entire genes ("one gene—one enzyme hypothesis"). These views are related to the genetic determinsim view of having a gene for every trait, in which every gene has a function. However, the primary functional units inside a cell are arguably its proteins. Their biochemical activities and biological roles depend crucially on their three-dimensional shapes and network organization, respectively.

One has also to take into account more abstract functional entities, i.e., network modules. These are also called 'structures' but do not refer to the shape of molecules. Its functions ought to be considered as Wouter's second notion (biological role), rather than biochemical activity. "Current 'systems' thinking attributes primary functional significance to the collective properties of molecular networks rather than to the individual properties of component molecules" (Shapiro, 2011, 129). "[A] discrete biological function can only rarely be attributed to an individual molecule [...]. In contrast, most biological functions arise from interactions among many components." (Hartwell et al., 1999, C47). Thus, we can attribute functions as biochemical activities to

individual molecules, wheres systems functions (biological roles) are attributed to organizational structures:

> "Finding a sequence motif (e.g., a kinase domain) in a new protein sheds light on its biochemical function; similarly, finding a network motif in a new network may help explain what systems-level function the network performs, and how it performs it." (Alon, 2003, 1867)

4 Does structure determine function?

Having distinguished between three notions of 'structure' and two notions of 'function', what about the statement 'structure determines function'? Is—in any of its different readings—a certain structure necessary or sufficient for a certain function?

The common textbook view according to Anfinsen has a clear answer: "the central dogma of structural biology is that a folded protein structure is necessary for biological function" (Wright and Dyson, 1999, 322). On first glance, it might appear plausible to assume that a particular structure (understood as molecular shape) is a necessary condition for the proper function of a biological structure (i.e., its biochemical activity). Loss of function is often associated with a loss of the three-dimensional shape of individual proteins. On the other hand, to go for the "sufficient" direction, changes in structures often lead to a decrease in functionality, up to a complete loss. Many diseases for which there are known molecular causes give support to this view. Often it is alterations in the sequence of DNA that result in changed protein shapes that lead to a functionality defect of the organism, which is the definition of a "molecular disease". Alterations of a protein's three-dimensional shape, however, do not necessarily lead to

loss of function. In many cases, changes are "silent", i.e., they don't cause any alteration in phenotype. In rare events, changes might even turn out to be "improvements", which is the driving force of evolutionary development.

However, evidence has been found in the recent years that a significant portion of proteins are intrinsically unstructured in order to be functional, see for example (Forman-Kay and Mittag, 2013). Does the discovery of intrinsically unstructured proteins challenge the relation between structure and function? "[D]isorder aficionados are calling for a complete reassessment of the structure-function paradigm" (Chouard, 2011, 151). Some protein domains fold only upon binding to a suitable target. Others, however, seem to never have an ordered state at all—they remain unstructured even in their functional state.

That a high similarity in sequence does not guarantee a similarity in structure or function has been shown by the Paracelsus Challenge: "a one-time prize of $1000, to be awarded to the first individual or group that successfully transforms one globular protein's conformation into another by changing no more than half the sequence" (Rose and Creamer, 1994, 3). One recent answer to this challenge resulted in the synthesis of two proteins which have 88% sequence identity but a different structure and a different function (Alexander et al., 2007).

Contrary to the view described above, the generalization that a stable three-dimensional structure is necessary or sufficient for a particular function does not hold. It remains true, however, that there is an intimate correlation between structure and function. Prediction tools based on this view are a powerful tool. An attempt to systematically predict the structure and function of proteins based on their amino acid sequence can be found, for example, in (Roy et al., 2010).

To complicate the picture, codon usage is also important: Zhou et al. (2013) have shown that the FRQ protein, which is involved in the circadian clock, is using non-optimal codons, thus translation speed is not optimal. After experimentally optimizing codon usage, the resulting protein—which has the exact same amino acid sequence—folds differently and is no longer functional. This shows that amino acid sequence by itself is not sufficient to determine the three-dimensional structure, let alone its function. In addition to the correct sequence, the folding process has to take place in a certain way which is influenced by the usage of codons and thus the availability of tRNAs, which influences the speed at which the ribosome can proceed translation. Usage of non-optimal codons gives the nascent polypeptide chain some time for the segments that have already been translated to fold in a certain conformation. If translation is too fast, certain intermediate folds which are necessary to reach the final functional conformation can be lost.

Another idea to keep in mind is that evolution operates pragmatically: structures are not the target of selection, functions are. Structures are being re-used for novel functions—there are many biological examples.

If structure does not *determine* function, if a particular structure (in any of its three meanings) is neither necessary nor sufficient for a particular function (in any of its two meanings), may there be another way in which structure and function are related? Perhaps there is a less stringent relationship? I will argue for a supervenience relation (McLaughlin and Bennett, 2018). But before developing this account, we need to clarify which notions of 'structure' and 'function' to use to capture actual scientific practice in molecular biology.

In order to speak about biological functions, a reglemented vocabulary is needed.

The most successful of these is gene ontology (GO) (Ashburner et al., 2000). Fascinating correlation analysis between three-dimensional protein structures from the Protein Data Bank (PDB) and GO terms can be found, for example, in (Hvidsten et al., 2009) and (Pal and Eisenberg, 2005).

According to the textbook picture, there is a linear chain of determination, leading from nucleotide sequences in the DNA via transcription to the nucleotide sequence of RNA, which leads via translation to the amino acid sequence of proteins. The sequence of amino acids, in turn, determines the three-dimensional structure of the protein, whose function, again, is determined by its structure. Given transitivity of this determination relation, one would only need to know the genomic sequence in order to have a complete picture ("blue print") of the functional organism. That is the "holy grail of molecular biology". And like the quest for the holy grail, it is doomed to fail. A strict determination relation does not even hold between the individual pairs.

The reason why the simplified scheme above is still part of the current research "paradigm" lies, on the one hand, in its scientific success: genomics and proteomics have provided unimaginable insights. On the other hand, it fits the mechanistic, reductionistic narrative that has been fashionable in molecular biology. Today, systems biology claims to provide a "holistic" alternative (Green, 2017).

But even without such a strict determination relation between structure and function, both concepts are central to explaining molecular mechanisms in research practice.

In order to understand why molecular biologists explain mechanisms with reference to structure and function, we need to understand what these concepts denote. In a first approximation, molecular biologists analyze a phenomenon by identifying its components that are responsible for the phenomenon in question. These components are the

structures that perform certain biochemical activities, which collectively bring about the phenomenon (biological role). The way in which these entities and their activities are organized is a different meaning of 'structure' which is as important in a mechanistic explanation as individual molecular structures are.

> "Despite the lack of an overarching theory, a Newtonian or quantum mechanics of its very own, molecular biology has become a unifying discipline in virtue of the powers of its techniques, its ability to extrapolate from the molecular to higher levels, and its synthesis of problems of form and function at the molecular level. This synthesis of form and function is a central, ill-understood, and historically important feature of molecular biology." (Burian, 1996, 68)

The ambiguity of the terms 'structure' and 'function' might be useful, for it can be applied to a broad variety of biological research strategies and activities. But, on the other hand, using the term same for different things causes confusion, and the use of metaphorical language might be obscuring certain features and difficulties with this approach.

More recent and thriving approaches in the life sciences have moved beyond the idea that there is a determination relation between structure and function and that by knowing the structure of a protein one could predict its biological function. Today's research in molecular biology is more centered around the *organizational structure* of biological mechanisms. In this way, the ambiguity of the term 'structure' suits to uphold the research slogan, since it can also be applied in a broader sense here than just molecular shapes. The organization of biological systems is the domain of the relatively

new discipline systems biology.

The three-dimensional shape is often a detail that does not contribute to the understanding of a mechanism, but to the contrary would only confuse the mechanistic picture which requires a certain level of abstraction in order to be comprehensive.

But still, how exactly do we get from molecular structures and their (structured) activities to biochemical activities and biological functions? That there might not exist a straightforward mapping from molecular shapes to their biochemical and biological function had been anticipated in the early days of molecular biology:

> "It [molecular biology] is concerned particularly with the *forms* of biological molecules, and with the evolution, exploitation and ramification of these forms in the ascent to higher and higher levels of organization. Molecular biology is predominantly three-dimensional and structural—which does not mean, however, that it is merely a refinement of morphology. It must of necessity enquire at the same time into genesis and function." (Astbury, 1952, 3, original emphasis)

Taking up Francis Crick's remark that "folding is simply a function of the order of the amino acids" (Crick 1958, 144), Morange comments that it is "obviously not a *simple* function" (Morange, 2006, 522). And he observes a semantic change in the meaning of 'function':

> "For Francis Crick, function meant the application of simple rules and principles. For specialists today, function is the result of a complex evolution [...] This shift in the meaning of a word is more than anecdotal. It reflects an active ongoing transformation of biology [...] The mechanistic models of

> molecular biology are no longer considered sufficient to explain the structures and functions of organisms. They have to be complemented and allied with evolutionary explanations" (Morange, 2006, 522).

In order to explain biological phenomena, there is no determination relation that would allow us to track everything down to the chemical and physical properties of proteins, let alone the nucleotide sequences of DNA. Of course, all these issues are relevant to the topic of reduction:

> "if [...] regulatory networks turn out to be crucial to explaining development (and evolution [...]), the reductionist interpretation *may* be in trouble. If network-based explanations are ubiquitous, it is quite likely that what will often bear the explanatory weight in such explanations is the topology of the network rather than the specific entities of which it is composed. [...] How topological an explanation is becomes a matter of degree: the more an explanation depends on individual properties of a vertex, the closer an explanation comes to traditional reduction. The components matter more than the structure. Conversely, the more an explanation is independent of individual properties of a vertex, the less reductionist it becomes." (Sarkar, 2008, 68, original emphasis)

5 Conclusion

Both terms, 'structure' and 'function', are highly ambiguous. So is the widely used conjunct phrase of 'structure and function' that is ubiquitous in biology, as well as the

even stonger claim 'structure determines function'. Perhaps this is why it can be used in many different contexts and for many different explanatory aims in biology. Although providing a certain framework of generality, I argue that a clarification of these concepts is beneficial—for conceptual and philosophical considerations, as well as for the way biologists think about the grand schemes like the "central dogma". Ideally, such an account would also have practical implications and benefit current biological research.

To sum up the results of my analysis, in molecular biology's explanatory practice, 'structure' may refer to:

1. the sequence of polymers,
2. the three-dimensional shape of molecules (or their parts), and
3. the way biological entities are organized.

Of course, different aspects of this distinction play different roles in the explanatory practice with respect to molecular mechanisms. The detailed shape of the interacting molecules is neither necessary nor sufficient for understanding its activities (although correlations are valuable prediction tools before doing experiments in the lab).

The ambiguity of the term 'function' depends on whether the explanation aims at answering the question how a mechanism works or how it came to work that way. Even in the first case one has to distinguish between:

1. the biochemical activity of individual components, and
2. the biological role of network structures.

Whereas biochemical activities of proteins can often be successfully predicted by homology modeling from known molecular shapes, the biological role is rarely an

intrinsic property of an isolated molecule. Rather, the biological role is the mechanistic result of an interaction network of several dynamically interacting molecules.

By comparing the combinatorial possibilities of the different meanings of 'structure' and 'function', a determination relation does not hold between any of them. Instead, I propose a supervenience relation: between the three-dimensional shapes of protein domains and their biochemical activities, and between interaction networks and their biological role. According to my analysis, this is what molecular biologist mean when they say 'structure determines function'.

References

Alexander, P. A., Y. He, Y. Chen, J. Orban, and P. N. Bryan (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences 104*(29), 11963–11968. doi:10.1073/pnas.0700922104.

Allen, C. (2009). Teleological notions in biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009 ed.). `http://plato.stanford.edu/archives/win2009/entries/teleology-biology/`.

Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science 301*(5641), 1866–1867. doi:10.1126/science.1089072.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science 181*(4096), 223–230. doi:10.1126/science.181.4096.223.

Anfinsen, C. B., E. Haber, M. Sela, and F. H. White, Jr (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences 47*(9), 1309–1314.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. *Nature Genetics 25*, 25–29. doi:10.1038/75556.

Astbury, W. T. (1952). Adventures in molecular biology. In *The Harvey Lectures. Delivered under the auspices of the Harvey Society of New York. 1950–51*, pp. 3–44. Charles C Thomas.

Burian, R. M. (1996). Underappreciated pathways toward molecular genetics as illustrated by Jean Brachet's cytochemical embryology. In S. Sarkar (Ed.), *The Philosophy and History of Molecular Biology: New Perspectives*, pp. 67–85. Kluwer Academic Publishers.

Chouard, T. (2011). Breaking the protein rules. *Nature 471*, 151–153. doi:10.1038/471151a.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy 72*(20), 741–765. doi:10.2307/2024640.

Forman-Kay, J. D. and T. Mittag (2013). From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure 21*(9), 1492–1499. doi:10.1016/j.str.2013.08.001.

Green, S. (2017). Philosophy of systems and synthetic biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 ed.). `https://plato.stanford.edu/archives/sum2017/entries/systems-synthetic-biology/`.

Gutte, B. and R. B. Merrifield (1971). The synthesis of ribonuclease A. *Journal of Biological Chemistry 246*, 1922–1941.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray (1999). From molecular to modular cell biology. *Nature 402*(6761 Suppl.), C47–C52. doi:10.1038/35011540.

Hvidsten, T. R., A. Lægreid, A. Kryshtafovych, G. Andersson, K. Fidelis, and J. Komorowski (2009). A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLoS ONE 4*(7), e6266. doi:10.1371/journal.pone.0006266.

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, and H. Wyckoff (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature 181*(4610), 662–666. doi:10.1038/181662a0.

McLaughlin, B. and K. Bennett (2018). Supervenience. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). `https://plato.stanford.edu/archives/spr2018/entries/supervenience/`.

Millikan, R. G. (1989). An ambiguity in the notion "function". *Biology and Philosophy 4*, 172–176. doi:10.1007/BF00127747.

Mitchell, S. D. and A. M. Gronenborn (2017). After fifty years, why are protein X-ray

crystallographers still in business? *The British Journal for the Philosophy of Science 68*(31), 703–723. doi:10.1093/bjps/axv051.

Morange, M. (2006). The protein side of the central dogma: Permanence and change. *History and Philosophy of the Life Sciences 28*(4), 513–524.

Pal, D. and D. Eisenberg (2005). Inference of protein function from protein structure. *Structure 13*, 121–130. doi:10.1016/j.str.2004.10.015.

Rose, G. D. and T. P. Creamer (1994). Protein folding: Predicting predicting. *PROTEINS: Structure, Function, and Genetics 19*, 1–3.

Roy, A., A. Kucukural, and Y. Zhang (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols 5*(4), 725–738. doi:10.1038/nprot.2010.5.

Sadowski, M. and D. T. Jones (2009). The sequence–structure relationship and protein function prediction. *Current Opinion in Structural Biology 19*, 357–362. doi:10.1016/j.sbi.2009.03.008.

Sarkar, S. (2008). Genomics, proteomics, and beyond. In S. Sarkar and A. Plutynski (Eds.), *A Companion to the Philosophy of Biology*, pp. 58–73. Blackwell Publishing Ltd.

Shapiro, J. A. (2011). *Evolution: a view from the 21st century.* FT Press Science.

Wouters, A. G. (2003). Four notions of biological function. *Studies in History and Philosophy of Biological and Biomedical Sciences 34*(4), 633–668. doi:10.1016/j.shpsc.2003.09.006.

Wright, L. (1973). Functions. *The Philosophical Review 82*(2), 139–168. doi:10.2307/2183766.

Wright, P. E. and H. J. Dyson (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology 293*, 321–331. doi:10.1006/jmbi.1999.3110.

Zhou, M., J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. Sachs, and Y. Liu (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature 495*, 111–115. doi:10.1038/nature11833.

Is Peer Review a Good Idea?*

Remco Heesen[†‡] Liam Kofi Bright[§]

September 19, 2018

Abstract

Pre-publication peer review should be abolished. We consider the effects that such a change will have on the social structure of science, paying particular attention to the changed incentive structure and the likely effects on the behavior of individual scientists. We evaluate these changes from the perspective of epistemic consequentialism. We find that where the effects of abolishing pre-publication peer review can be evaluated with a reasonable level of confidence based on presently available evidence, they are either positive or neutral. We conclude that on present evidence abolishing peer review weakly dominates the status quo.

*Both authors contributed equally. Thanks to Justin Bruner, Adrian Currie, Cailin O'Connor, and Jan-Willem Romeijn for valuable comments. RH was supported by an Early Career Fellowship from the Leverhulme Trust and the Isaac Newton Trust. LKB was supported by NSF grant SES 1254291.

[†]Department of Philosophy, School of Humanities, University of Western Australia, Crawley, WA 6009, Australia. Email: remco.heesen@uwa.edu.au.

[‡]Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK.

[§]Department of Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: liamkbright@gmail.com.

1 Introduction

Peer review plays a central role in contemporary academic life. It sits at the critical juncture where scientific work is accepted for publication or rejected. This is particularly clear when the results of scientific work are communicated to non-scientists, e.g., by journalists. The question "Has this been peer reviewed?" is commonly asked, and a positive answer is frequently taken to be a necessary and sufficient condition for the results to be considered serious science.

Given these circumstances, one might expect peer review to be an important topic in the philosophy of science as well. Peer review should arguably play a more prominent role in the debate about demarcation criteria (what separates science from other human pursuits?), as it seems to be used in practice exactly to differentiate scientific knowledge from other claims to knowledge, at least by journalists. Yet as far as we know, social-procedural accounts of science, like the one found in Longino (1990), remain in the minority and usually do not place great emphasis on peer review in particular. Aside from this particular debate, there are normative questions about the proper epistemic role of peer review and more practical questions about the extent to which it manages to fulfill them, all of which should interest philosophers of science.

But there has been surprisingly little work on peer review by philosophers of science. Most of what exists has focused on the role of biases in peer review, see for example Saul (2013, §2.1), Lee et al. (2013), Jukola (2017), Katzav and Vaesen (2017), and Heesen (2018). We are not aware of any philosophical discussion of the strengths and weaknesses of peer review as such (the above examples presuppose its overall legitimacy by discussing its implementation). Some work along these lines does exist outside of philosophy, either in the form of opinion pieces (Gowers 2017) or occasionally full-length articles (Smith 2006). Such work tends to be vague about the normative standard against which peer review or its alternatives are to be

evaluated, something we aim to remedy in section 2.

Here we bring together the work of philosophers of science (especially social epistemologists of science) who have written about the strengths and weaknesses of various aspects of the social structure of science and empirical work about the effects of peer review. We argue that where philosophers of science have claimed the social structure of science works well, their arguments tend to rely on things other than peer review, and that where specific benefits have been claimed for peer review, empirical research has so far failed to bear these out. Comparing this to the downsides of peer review, most prominently the massive amount of time and resources tied up in it, we conclude that we might be better off abolishing peer review.

Some brief clarifications. Our target is pre-publication peer review, that is the review of a manuscript intended for publication, where publication is withheld until one or more editors deem the manuscript to have successfully passed peer review. We set aside other uses of peer review (e.g., of grant proposals or conference abstracts) and we explicitly leave room for post-publication peer review, where manuscripts are published before review. Because of this last point, some readers may think that our terminology ('abolishing pre-publication peer review') suggests a more dramatic change than what we actually advocate. We invite such readers to substitute in their preferred terminology. We should also clarify that we use 'science' in a broad sense to include the natural sciences, the social sciences, and the humanities.

The overall structure of our argument is as follows. We think there are a number of clear benefits to abolishing pre-publication peer review. In contrast, while various benefits of the existing system (downsides of abolishing peer review) have been suggested, we do not think there exist any that have clear empirical support. Insofar as empirical research exists, it is ambiguous in some cases, and speaks relatively clearly against the claimed benefit of the existing system in others. While we admit to a number of cases where the evidence is ambiguous or simply lacking (see especially section 5), we claim

that the present state of the evidence suggests that abolishing pre-publication peer review would lead to a Pareto improvement: each factor considered is either neutral or favors our proposal.

Our primary aim here is to evaluate the current system, but we believe that is only really possible by comparing it to an alternative. We are not claiming that the proposal we put forward is the best of all possible alternatives. It has been constructed to be a system which could constitute a Pareto improvement over the current system. Given that it has not actually been implemented yet, we cannot guarantee it would work as advertised or what empirical properties it would have. But in offering a relatively specific alternative, we hope to get people thinking about real change, which pointing out problems with the present system has so far failed to do.

Even despite this proviso, we realize that ours is a strong claim, and our proposal a large change to the social structure of science. It is therefore important to highlight that our central claim concerns the balance of presently available evidence. We are not further claiming that the matter is so conclusively settled as to render further research superfluous or wasteful. On the contrary, we think there are a number of points in our argument where the presently available evidence is severely limited, and we take the calls for further empirical research we make in those places to be just as important a part of the upshot of our paper as our positive proposal. We hope, therefore, that even a skeptical reader will read on; if not to be convinced of the need of abolishing pre-publication peer review, then at least to see where in our view their future research efforts should concentrate if they are to shore up pre-publication peer review's claims to good epistemic standing.

2 Setting the Stage

The purpose of peer review is usually construed in terms of quality control. For example, Katzav and Vaesen (2017, 6) write "The epistemic role of peer

review is assessing the quality of research", and this seems to be a common sentiment per, e.g., Eisenhart (2002, 241) and Jukola (2017, 125). But how well does peer review succeed in its purpose of quality control? The empirical evidence (reviewed below) is mixed at best. As one prominent critic puts it, "we have little evidence on the effectiveness of peer review, but we have considerable evidence on its defects" (Smith 2006, 179).

Peer review's limited effectiveness would perhaps not be a problem if it required little time and effort from scientists. But in fact the opposite is true. Going from a manuscript to a published paper involves many hours of reviewing work by the assigned peer reviewers and a significant time investment from the editor handling the submission. The editor and reviewers are all scientists themselves, so the epistemic opportunity cost of their reviewing work is significant: instead of reviewing, they could be doing more science.

Given these two facts—high (epistemic) costs and unclear benefits—we raise the question whether it might be better to abolish pre-publication peer review. In the following we provide our own survey and assessment of the evidence that bears on this question. Our conclusions are not sympathetic to peer review. However, we encourage any proponents of peer review to give their own assessment. We only ask that any benefits claimed for peer review are backed up by empirical research, and that they are epistemic benefits, i.e., we ask for empirical evidence that peer review makes for better science on science's own terms.

We take the status quo to be as follows. The vast majority of scientific work is shared through journal publications, and the vast majority of journals uses some form of pre-publication peer review. Ordinarily this means that an editor assigns one to three peers (scientists whose expertise intersects the topic of the submission), who provide a report and/or verdict on the submission's suitability for publication. The peer reviews feed into the final judgment: the submission is accepted or rejected with or without revisions.

Our proposal is to abolish pre-publication peer review. Scientists them-

selves will decide when their work is ready for sharing. When this happens, they publish their work online on something that looks like a preprint archive (although the term "preprint" would not be appropriate under our proposal). Authors can subsequently publish updated versions that reply to questions and comments from other scientists, which may have been provided publicly or privately. Most journals will probably cease to exist, but the business of those that continue will be to create curated collections of previously published articles. Their process for creating these collections will presumably still involve peer review, but now of the post-publication variety.

Our proposal is in line with how certain parts of mathematics and physics already work: uploading a paper to arXiv is considered publishing it for most purposes, with journal peer review and publication happening almost as an afterthought (Gowers 2017). Indeed, journal publication can function as something like a prize, accruing glory to the scientist who achieves it but doing little to actually help spread or diffuse the idea beyond calling attention to something that has already been made public elsewhere. We are not aware of any detailed comparative studies of the effects these changes have had in those fields, so we will not rest any significant part of our argument on this case. But for those who worry that science will immediately and irrevocably fall apart without peer review, we point out that this does not appear to have happened in the relevant parts of mathematics and physics.

In the remainder of this paper we break down the consequences of our proposal. Our strategy here is to focus on a large number (hopefully all) aspects of the social structure of science that will be affected. In particular, the reader may already have a particular objection against our proposal in mind. We encourage such a reader to skip ahead to the section where this objection is discussed before reading the rest of the paper.

For example, one reader may think that peer review as currently practiced is important because it forces scientists to read and review each other's work, and without peer review they will spend less time on such tasks. This

is discussed in section 3.2. Another reader may worry that without peer review and the journal publications that go with them it will be more difficult to evaluate scientists for hiring or promotion (section 3.5). Yet another reader may be concerned about losing peer review's ability to prevent work of little merit from being published, or at least to sort papers into journals by epistemic merit so scientists can easily find good work (section 4.1). A fourth reader might think peer review plays an important role in detecting fraud or other scientific malpractice (section 4.2). A fifth reader may think the guarantee provided to outsiders when something has been peer reviewed is an important reason to preserve the status quo (section 5.1). And a sixth reader may want to point out that anonymized peer review gives relatively unknown scientists a chance at an audience by publishing in a prestigious journal, whereas on our proposal perhaps only antecedently prominent scientists will have their work read and engaged with (section 5.2).

Other aspects of the social structure of science that will be considered: whether and when scientists share their work (section 3.1), how many papers are published by women or men (section 3.3), library resources (section 3.4), the power of editors as gatekeepers (section 3.6), science's susceptibility to fads and fashions (section 4.3), and ways to get credit for scientific work other than through journal publications (section 4.4). In each case we evaluate whether the net effects of our proposal on that aspect can be expected to be positive. To tip our hand: aspects where we will claim a benefit are gathered in section 3, aspects where we expect little or no change are in section 4, and aspects that we consider neutral due to a present lack of evidence are in section 5.

In making these evaluations, we commit to a kind of epistemic consequentialism (cf. Goldman 1999). One may think of what we are doing as roughly analogous to the utilitarian principle, where for each issue our yardstick is whether pre-publication peer review shall generate the greatest amount of knowledge produced in the least amount of time. More specifically, we con-

sider changes in the incentive structure and expected behaviour of scientists, as well as other changes that would result from abolishing pre-publication peer review. We evaluate these changes in terms of their expected effect on the ability of the scientific community to produce scientific knowledge in an efficient manner. Working out in detail what such an epistemic consequentialism would look like would be very complicated, and we do not attempt the task here. For most of the issues we consider, we think that the calculus is sufficiently clear that fine details do not matter. Where it is unclear (the issues discussed in section 5) we think this results from ignorance of empirical facts about the likely effect of policies, rather than conceptual unclarity in the evaluative metric. So we do not need to use our consequentialist yardstick to settle any difficult tradeoffs. All we need for our purposes is to make it clear that we are evaluating the peer review system by how well it does in incentivizing efficient knowledge production.

What do we mean by the incentive structure of science, mentioned in the previous paragraph? This addresses the motivations of scientists. Scientists are rewarded for their contributions with credit, i.e., with recognition from their peers as expressed through such things as awards, citations, and prestigious publications (Merton 1957, Hull 1988, Zollman 2018). Scientific careers are largely built on the reputations scientists acquire in this way (Latour and Woolgar 1986, chapter 5). As a result, scientists engage in behaviors that improve their chances of credit (Merton 1969, Dasgupta and David 1994, Zollman 2018).

While individual scientists may be motivated by credit to different degrees (curiosity, the thrill of discovery, and philanthropic goals are important motivations for many as well), the effect on careers means that credit-maximizing behavior is to some extent selected for. Thus we think it important to ensure that our proposal does not negatively affect the incentives currently in place for scientists to work effectively and efficiently.

3 Benefits of Abolishing Peer Review

3.1 Sharing Scientific Results

An important feature of (academic) science is that there is a norm of sharing one's findings with the scientific community. This has been referred to as the communist norm (Merton 1942). In recent surveys, scientists by and large confirm both the normative force of the communist norm and their actual compliance (Louis et al. 2002, Macfarlane and Cheng 2008, Anderson et al. 2010). This norm is epistemically beneficial to the scientific community, as it prevents scientists from needlessly duplicating each other's work.

Will abolishing peer review affect this practice? In order to answer this question, we need to know what motivates scientists to comply with the communist norm, that is to share their work. On the one hand there is the feeling that they ought to share generated by the existence of the norm itself. There is no reason to expect this to be changed by abolishing peer review.

On the other hand there is the motivation generated by the desire for credit. According to the priority rule, the first scientist to publish a particular discovery gets the credit for it (Merton 1957, Dasgupta and David 1994, Strevens 2003). So a scientist who wants to get credit for her discoveries has an incentive to publish them as quickly as possible, in order to maximize her chances of being first. Recent work suggests that this applies even in the case of smaller, intermediate discoveries (Boyer 2014, Strevens 2017, Heesen 2017b). All of this helps motivate scientists to share their work.

If peer review were to be abolished, the communist norm and the priority rule would still be in effect, so scientists would still be motivated to share their work as quickly as possible. However, the following change would occur.

In the absence of pre-publication peer review, scientists would be able to share their discoveries more quickly. In the current system, peer review can hold up publication for significant amounts of time, especially in the case of fields with high rejection rates or long turnaround times. During this time,

other scientists cannot build on the work and may spend their time needlessly duplicating the work. Cutting out this lag by letting scientists publish their own work when they think it is ready will speed up scientific progress. While being faster is not always better (it may increase the risk of error, cf. Heesen 2017c), in this case delays in publication are reduced without any reduction in the time spent on the scientific work itself.

To some extent this already occurs. Scientists, especially well-connected scientists, already share preprints that make the community aware of their work in advance of publication. For people who regularly do this, practically speaking little would change upon adopting the system we advocate. However, our proposal turns pre-journal-publication dispersal of work from a privilege of a well-connected few into the norm for everyone.

On this point, then, abolishing peer review is a net positive, as scientists will still be incentivized to share their work as soon as possible, but the delays associated with pre-publication peer review are removed.

3.2 Time Allocation

The current system restricts the way scientists are allowed to spend their time. For each paper submitted to a journal, a number of scientists are conscripted into reviewing it, and at least one editor has to spend time on that paper as well.

On our proposal, scientists would be free to choose how much of their time to spend reading and reviewing others' work as compared to other scientific activities. Some scientists would spend less time reviewing, some scientists would spend more, and some would spend exactly as much as under the current system.

For scientists in the latter category our proposal makes no difference, while for scientists in the other two categories our proposal represents a net improvement of how they spend their time, at least in their own judgments. We think people are the best judges of how to use their own time and labor.

We thus trust scientists' decisions in these regards, and welcome changes that would render many scientists' choices about how to allocate their own labor independent of the preferences of the relatively small number of editorial gatekeepers.

So we assume that scientists are well-placed to judge how best to use their own abilities to meet the community's epistemic needs. We claim, moreover, that the reward structure of science is set up so as to make it in their interest to do so: the credit economy incentivizes scientists to spend their time on pursuits the epistemic value of which will be recognized by the community (Zollman 2018). Hence freeing up the way scientists allocate their time leads to net epistemic benefits to the scientific community.

One might object that journals perform a useful epistemic sorting role, telling scientists what is worth spending their time on. We will address these concerns in section 4.1.

One might think that this would lead scientists to spend significantly less time reading and reviewing others' work. If this is right, we still think it would be an overall improvement for the reasons mentioned above. But we also want to point out that this is not as obvious a consequence as it may seem. Here are two reasons to expect scientists to spend as much time or more reading and reviewing on our proposal. First, for many scientists reading and reviewing are intrinsically valuable and can help their own research. Second, the current system provides no particular incentive to read and review either: scientists agree to review only because they independently want to or because they feel an obligation to the research community. While no one scientist is conscripted, at the group level editors are going to keep going until they find someone. This can amount to picking whomever is most weak-willed or under some extra-epistemic social pressure. It is not obvious that this way of deciding who does the reviewing has much to recommend it. Any rewards that exist for reviewing will still exist on our proposal, and may be amplified by the possibility of making post-publication reviews public.

3.3 Gender Skew in Publications

Male scientists publish more, on average, than female scientists, a phenomenon known as the productivity puzzle or productivity gap (Zuckerman and Cole 1975, Valian 1999, Prpić 2002, Etzkowitz et al. 2008). Several explanations have been suggested, none of which are entirely satisfactory (see especially Etzkowitz et al. 2008, 409–412). Two of these explanations that are relevant to our concerns here are the direct effects of gender bias and the indirect effects of the expectation of gender bias.

There is some evidence of gender bias in peer review, although this is not unambiguous (see Lee et al. 2013, 7–8, Lee 2016, and references therein). Insofar as there is gender bias—in the sense of women's work being judged more negatively by peer reviewers—abolishing peer review will remove this and help level the playing field for men and women. We expect positive epistemic consequences from the removal of these arbitrarily different standards.

While the evidence of gender bias in peer review is not entirely clear-cut, there is good evidence that women *expect* to face gender bias in peer review (see Lee 2016, Bright 2017b, Hengel 2018, and references therein). In an effort to overcome this perceived bias, women tend to hold their own work to higher standards. Hengel (2018) provides evidence that women spend more time correcting stylistic aspects of their paper during peer review, presumably due to higher expectations of scrutiny on such apparently superficial elements of their work. On the plausible assumption that if women have higher standards for each paper they will produce fewer papers overall, this means that the mere expectation of gender bias can contribute to the productivity gap.

After abolishing peer review both women and men will hold their work primarily to their own individual standards of quality, and secondarily to their expectations of the response of the entire scientific community, but not to their expectations of the opinion of a small arbitrary group of gatekeepers. We do not know whether this will lead the women to behave more like the men (producing more papers) or the men to behave more like the women

(holding individual papers to a higher standard of quality). However, in line with our view above that scientists are well-placed to judge how best to spend their own time, we take it that any resulting change in behavior will be a net epistemic positive.

3.4 Library Resources

Journal subscription fees currently take up a large amount of library resources (RIN 2008, Van Noorden 2013). To summarize some key figures from the 2008 report: research libraries in the UK spent between £208,000 and £1,386,000 on journal subscriptions annually (and that was a decade ago, with subscriptions having risen substantially since). The cost for publishing and distributing a paper was estimated to be about £4,000, or about £6.4 billion per year in total. Savings from moving to author-paid open access were estimated at £561 million, about half of which would directly benefit libraries.

On our proposal, this is replaced by the cost of maintaining one or more online archives of scientific publications. The example of existing large preprint archives like arXiv and bioRxiv suggests that maintaining such archives can be done at a fraction of the cost currently spent on journal subscription fees. As a rough guideline, Van Noorden (2013) estimates maintenance costs of arXiv at just $10 per article. So our proposal involves significant savings on library resources, which could be used to expand collections, retain more or better trained staff, or other purposes that would be of epistemic benefit to the scientific community.

Two additional effects should be considered in relation to this. First, the fact that the online archive will be open access means that scientific publications will be available to everyone, not just to those with a library subscription or some other form of access to for-profit scientific journals.

Second, the fact that any value added by for-profit journals would be taken away. The two tasks currently carried out by journals that could

plausibly be supposed to add value to scientific publications are peer review and copy-editing (Van Noorden 2013). It is the purpose of all other sections of this paper to argue that peer review does not in fact (provably) add value, so we set that aside. This leaves copy-editing. We propose that libraries use some of the funds freed up from journal subscriptions to employ some copy-editors. Each university library would make copy-editing services available to the scientists employed at that university. We contend that, after paying for the maintenance of an online archive and a team of copy-editors, under our proposal libraries would still end up with more resources for other pursuits than under the current system.

We note that this particular advantage of our proposal is a bit more historically contingent than the others. There seems to be no particular reason why pre-publication peer review has to be implemented through for-profit journals, and if the open access movement has its way we might be able to free up these library resources without abolishing pre-publication peer review. But our proposal also achieves this goal, and so we count it as an advantage relative to the system as it is currently actually implemented.

3.5 Scientific Careers

The 'publish or perish' culture in science has been widely noted (e.g., Fanelli 2010). Universities judge the research productivity of scientists through their publications in (peer reviewed) journals, with some focusing more on 'quantity' (counting publications) and others on 'quality' (publishing in prestigious journals). Scientific journals and the system of pre-publication peer review thus play an important role in shaping scientific careers. What will become of this if peer review is abolished?

We note first that the 'publish or perish' culture is a subset of a larger system which we discussed above: the credit economy. Publishing in a journal is one way to receive credit for one's work, but there are others, most prominently citations and awards. Scientific careers depend on all of these,

with different institutions weighting quantity of publications, quality of publications, citation metrics, and awards and other honors differently.

Any of these types of credit represents some kind of recognition of the scholarly contributions of the scientist by her peers. But here we distinguish two types of credit, which we will call short-run credit and long-run credit. Getting a paper through peer review yields a certain amount of credit: more for more prestigious journals, less for less prestigious ones. But this is short-run credit in the following sense. The editor and the peer reviewers judge the technical adequacy and the potential impact of the paper, shortly after it is written. Their judgment is essentially a prediction of how much uptake the paper is likely to receive in the scientific community.

In contrast, citations (as well as awards, prizes, inclusion in anthologies or textbooks, etc.) represent long-run credit. They *are* the uptake the paper receives in the scientific community. Long-run credit is both a more considered opinion of the scientific importance of the paper and a more democratic one (citations can be made by anyone, and awards usually reflect a consensus in the scientific community, whereas peer review is normally done by up to three individuals). So long-run credit reflects a more direct and better estimate of the real epistemic value of a contribution to science.

So what would the effect of our proposal be? For better or worse, our proposal does not make it impossible for universities to use metrics to judge research productivity. While journal rankings and impact factors would disappear, citation metrics for individual scientists and papers would still be available. This may mean that universities stop judging their scientists based on the impact factors of the journals they publish in and start judging them on the actual citation impact of their papers. More generally, our proposal will decrease or remove the role of short-run credit in shaping career outcomes and increase the role of long-run credit, which we take to be a better measure of scientific importance. So we think this is an improvement on the status quo.

What about junior hires and related career decisions, where long-run credit may be absent or minimal? If abolishing peer review means completely getting rid of journals and the associated prestige rankings, this robs hiring departments of some information regarding the scientific importance of candidates' work. If this means those on the hiring side need to read and form an opinion of candidates' work for themselves, we do not think that is a bad thing. This would of course take time, but if journals and peer review are completely abolished, that just means the time spent reviewing the paper is transferred to the people considering hiring the scientist, which again, we do not think is a bad thing. In fact, since very few academics are on a hiring committee year after year, whereas referee requests are a constant feature while one is in the community, we think that even this added burden when hiring might still be a net time-saver for academics.

But it does not have to be that way. We never said journals and peer review have to be completely abolished—our proposal in section 2 explicitly suggests journal issues may still appear, but as curated collections of articles based on post-publication peer review. So short-run credit based on journal prestige need not disappear. It need not even be slower as there is no particular reason post-publication peer review needs to take longer than pre-publication peer review. But there is the added advantage that the paper is already published while it undergoes peer review, so the wider community outside the assigned reviewers also has a chance to respond before it is included in a journal.

3.6 The Power of Gatekeepers

The discussion immediately above touched on another effect, one that we think is worth bringing out as a benefit of our proposal in its own right. As mentioned our proposal suggests that in evaluating the importance of scientific work we decrease our reliance on short-run credit (journal prestige), with a corresponding increase in long-run credit (citations, among other things).

This means that the overall credit associated with a particular paper depends less on the judgments made by an editor and a small number of reviewers, and more on its actual uptake in the larger scientific community.

Editors in particular currently play a large role in determining which scientific work is worthy of attention, as they are a relatively small group of people with a deciding vote in the peer review process of a large number of papers. They are often referred to as gatekeepers for this reason (Crane 1967). Our proposal entails significantly decreasing both the prevalence and importance of this role. By replacing some of this importance with long-run credit, which comes from the scientific community as a whole, it makes the evaluation of scientific work a more democratic process. Not only is there some reason to think that democratic evaluation of scientific claims is more in line with general communal norms accepted within science (Bright et al. 2018), but general arguments from democratic theory and social epistemology of science give epistemic reason to welcome the increased independence of judgment and evaluation this would introduce (List and Goodin 2001, Heesen et al. forthcoming, Perović et al. 2016, 103–104).

4 Where Peer Review Makes No Difference

In this section we consider a number of aspects of the scientific incentive structure for which we think a case can be made that abolishing peer review will leave them basically unaffected. This serves partially to forestall objections to our proposal that we anticipate from defenders of the peer review system, and partially to avoid overstating our case—in some of what follows we argue that abolishing peer review will likely have no effect in cases where one might have expected it to be beneficial.

4.1 Epistemic Sorting

Given the stated purpose of peer review mentioned in section 2 the first and most apparent disadvantage of our proposal is that it would remove the epistemic filter on what enters into the scientific literature. One might worry that the scientific community would lose the ability to maintain its own epistemic standards, and thus the general quality of scientific research would be reduced. We argue here that despite the intuitive support this idea might have, the present state of the literature on scientific peer review does not support it.

Separate out two kind of epistemic standards one may hope that the peer review system maintains. First, that peer review allows us to identify especially meritorious work and place it in high profile journals, while ensuring that especially shoddy work is kept from being published. Call this the 'epistemic sorting' function of peer review. Second, that peer review allows for the early detection of fraudulent work or work that otherwise involves research misconduct. Call this the 'malpractice detection' function of peer review. We deal with each of these in turn.

Let us step back and ask why, from the point of view of epistemic consequentialism, one would want peer review to do any sort of epistemic sorting. We take the answer to be that epistemic sorting helps scientists fruitfully direct their time and energy by selecting the best work and bringing it to scientists' attention through publication in journals. They read and respond to that which is most likely to help them advance knowledge in their field.

How could peer review achieve this? One might hope that peer review functions by keeping bad manuscripts out of the published literature and letting good manuscripts in. This, however, is a non-starter. There are far too many journals publishing far too many things, with standards of publication varying far too wildly between them, for the sheer fact of having passed peer review somewhere to be all that informative as to the quality of a manuscript.

Instead, if peer review is to serve anything like this purpose it must be because reviewers are able (even if imperfectly) to discern the relative degree of scientific merit of a work, and sort it into an appropriately prestigious journal. Epistemic sorting happens not via the binary act of granting or withholding publication, but rather through sorting manuscripts into journals located on a prestige hierarchy that tracks scientific merit.

A necessary condition for epistemic sorting to work as advertised is that reviewers be reliable guides to the merit of the scientific work they review. Our first critique is that this necessary condition does not seem to be met. Investigation into reviewing practices has not generally found much inter-reviewer reliability in their evaluations (Peters and Ceci 1982, Ernst et al. 1993, Lee et al. 2013, 5–6). What this means is that one generally cannot predict what one reviewer will think of a manuscript by seeing what another reviewer thought. If there was some underlying epistemic merit scientists were accurately (even if falteringly) discerning by means of their reviews, one would expect there to be correlations in reviewers evaluations. However, this is not what we find. Indeed, one study of a top medical journal even found that "reviewers. . . agreed on the disposition of manuscripts at a rate barely exceeding what would be expected by chance" (Kravitz et al. 2010, 3). Findings like these are typical in the literature that looks at inter-reviewer reliability (for a review of the literature see Bornmann 2011, 207). The available evidence does not provide much support for the idea that pre-publication peer review detects the presence of some underlying quality.

Our second critique of the epistemic sorting idea speaks more directly to the ideal it tracks. We are not persuaded that the best way to direct scientists' attention is to continually alert them to the best pieces of individual work, and have them proportion their attention according to position on a prestige hierarchy. We take it the intuition behind this is a broadly meritocratic one. This intuition has been challenged by some modeling work (Zollman 2009). While Zollman retained some role for peer review, his model

still found that striving to select the best work for publication is not necessarily best from the perspective of an epistemic community; his model favored a greater degree of randomization.

We do not wish to rest our case on the results of one model which in any case does not fully align with our argument, but it highlights that the ideal of meritocracy stands in need of more defense than it is typically given. We take it that scientists most fruitfully direct their attention to that package of previous work and results which, when combined, provides them with the sort of information and perspectives they need to best advance their own epistemically valuable projects. It is a presently undefended assumption that this package of work should be composed of works which are themselves individually the most meritorious work, or that paying attention to the prestige hierarchy of journals and proportioning one's attention accordingly will be useful in constructing such a package. Hence, even if it did turn out that the peer review system could sort according to scientific merit, it is an underappreciated but important fact that this is not the end of the argument. Further defense of the purpose of this kind of epistemic sorting is needed from the point of view of epistemic consequentialism.

Before moving on we note a potential objection. Even if one did not think that peer review was detecting some underlying quality or interestingness, one might think that the process of feedback and revision which forms part of the peer review system would be beneficial to the epistemic quality of the scientific literature. In this way epistemic sorting may have a positive epistemic effect even if it fails in its primary task.

However, this returns us to the points regarding gatekeepers and time allocation from section 3. We are not opposed to scientists reading each other's work, offering feedback, and updating their work in light of that. This can indeed lead to improvements (Bornmann 2011, 203), though in this context it is worth noting the results of an experiment in the biomedical sciences, which found that attempting to attach the allure of greater prestige

to more epistemically high caliber work did little to actually improve the quality of published literature (Lee 2013). Fully interpreting these results would require discussion of the measures of quality used in such literature. We do not intend to do that here, since we do not intend to dispute the point that it is desirable for scientists to give feedback and respond to it.

We would expect this sort of peer-to-peer feedback to continue under a system without pre-publication peer review. Curiosity, informal networking, collegial responsibilities, and the credit incentives to engage with others' work and make use of new knowledge before others do; these would all be retained even without pre-publication peer review. What would be eliminated is the assignment of reviewing duties to papers that scientists did not independently decide were worth their time and attention, and the necessity of giving uptake to criticism (in order to publish) independently of an author's own assessment of the value of that feedback.

We thus conclude that, from the point of view of epistemic consequentialism, there is presently little reason to believe that a loss of the epistemic sorting function of pre-publication peer review would be a loss to science. Inclusion in the literature does not do much to vouch for the quality of a paper; the evidence does not favor the hypothesis that reviewers are selecting for some latent epistemic quality in order to sort into appropriate journals; and the ideal underlying the claimed benefits of epistemic sorting is dubious. While peer reviewers do give potentially valuable feedback, there is no particular reason to think that changes in how scientists decide to spend their time would make things worse in this regard, and (per our arguments in section 3) some reason to think that they would make things better.

4.2 Malpractice Detection

The other way peer review might uphold epistemic standards is through malpractice detection. However, once again, the literature does not support this. A number of prominent cases of fraudulent research managed to sail

through peer review. Upon investigation into the behavior of those involved it was found there was no reason to think that peer reviewers or editors were especially negligent in their duties (Grant 2002, 3). Peer reviewers report unwillingness to challenge something as fraudulent even where they have some suspicion that this is so, and avoid the charge (Francis 1989, 11–12). A criminologist who looked into fraudulent behavior in science reported that "virtually no fraudulent procedures have been detected by referees because reading a paper is neither a replication nor a lie-detecting device" (Ben-Yehuda 1986, 6). A more recent survey of the evidence found, at the least, no consistent pattern in journals' self-reported ability to detect and weed out fraudulent results (Anderson et al. 2013, 235).

Even if the prospect of peer review puts some people off committing fraud, the fact that it is so unreliable at detecting fraud suggests that this is a very fragile deterrence system indeed. Even this psychological deterrence would be rapidly undermined by more adventurous souls, or those pushed by desperation, since many would quickly learn that pre-publication peer review is a paper tiger.

Conversely, there are various ways for malpractice detection to operate in the absence of peer review. These include motive modification (Nosek et al. 2012, Bright 2017a), encouraging post-publication replication and scrutiny (Bruner 2013, Romero 2017), and the sterner inculcation of the norms of science coupled with greater expectation of oversight among coworkers (Braxton 1990). All of these methods of deterring fraud or meliorating its effects would still be available under our proposal.

What evidence we now have gives little reason to suppose that abolishing pre-publication peer review is any great loss to malpractice detection. Thus in this regard our proposal would make no great difference to the epistemic health of science. Combining this with the discussion of epistemic sorting, we conclude there is presently no reason to believe pre-publication peer review is adding much value to science by upholding epistemic standards.

4.3 Herding Behavior

Where above we argued that pre-publication peer review is not making a positive difference often claimed for it, in this section we downplay a potential benefit of our proposal. A consistent worry about scientific behavior is that it is subject to fads or, in any case, some sort of undesirable herding behavior (see, e.g. Chargaff 1976, Abrahamson 2009, Strevens 2013). A natural thought is that pre-publication peer review encourages this, since by its nature it means that to get new ideas out there one must convince one's peers that the work is impressive and interesting. It has thus been claimed that pre-publication peer review encourages unambitious within-paradigm work that unduly limits the range of scientific activity (Francis 1989, 12). Reducing the incentive to herd might thus be claimed as a potential benefit of our proposal. However, we are not convinced that it is pre-publication peer review that is doing the harmful work here.

As mentioned above, our proposal eliminates or significantly reduces the importance of short-run credit, the credit that accrues to one in virtue of publishing in a (more or less prestigious) scientific journal. Long-run credit, on the other hand, is left untouched. Under any sort of credit system, a scientist needs to do work that the community will pay attention to, build upon, and recognize her for. The mere fact that (she believes that) her peers are interested in a topic and liable to respond to it is thus still positive reason to adopt a topic. This is true even if the scientist would not judge that topic to be the best use of her time if she were (hypothetically) free from the social pressures and constraints of the scientific credit system.

The best that could be said about our proposal in this regard is that scientists would not specifically have to pass a jury of peers before getting their work out there. But given that we anticipate continued competition for the attention of scientific coworkers, it is hard to say what the net effect in encouraging more experimental or less conformist scientific work would be.

Whatever conformist effects the credit incentive has (see also the discus-

sion immediately below) do not depend on whether it is short- or long-run credit one seeks. The conformism comes from the fact that credit incentives focus scientists' attention on the predicted reaction of their fellow scientists to their work. Pre-publication peer review might make this fact especially salient by bringing manuscripts before a jury of peers before they may be entered into the literature. But even without pre-publication peer review the credit-seeking scientist must be focused on her peers' opinions. So there is no particular reason to think that removing the pre-publication scrutiny of manuscripts will free scientists from their own anticipations of the fads and fashions of their day.

4.4 Long-Run Credit

We end this section by noting that many of the effects of the credit economy of science studied by social epistemologists really concern long-run credit rather than the short-run credit affected by retaining or eliminating pre-publication peer review. This point is not restricted to herding behavior.

For instance, social epistemologists have studied both the incentive to collaborate, and various iniquities that can arise when scientists do not start with equal power when deciding who shall do what work and how they shall be credited (Harding 1995, Boyer-Kassem and Imbert 2015, Bruner and O'Connor 2017, O'Connor and Bruner forthcoming). Whether or not manuscripts would have to pass pre-publication peer review in order to enter the scientific literature, there would still be benefits in the long run to collaboration, and (alas) there would still be social inequalities that allow for iniquities to manifest in the scientific prestige hierarchy.

For another example, social epistemologists have studied the ways in which the credit incentive encourages different strategies for developing a research profile or molding one's scientific personality to be more or less risk-taking (Weisberg and Muldoon 2009, Alexander et al. 2015, Thoma 2015). Once again, pre-publication peer review plays no particular role in the analy-

sis. The incentives to differentiate oneself from one's peers (without straying too far from the beaten path) and to mold one's personality accordingly exist independently of pre-publication peer review.

Two especially influential streams of work in the social epistemology of science have been the study of the division of cognitive labor (Kitcher 1990, Strevens 2003), and the role of credit in providing a spur to work in situations with a risk of under-production (Dasgupta and David 1994, Stephan 1996). These two streams have directed the focus of the field, and have formed some of the chief defenses of the credit economy of science as it now stands (but see Zollman 2018, for a more critical take).

We mention them here because pre-publication peer review or short-run credit again plays no particular role in the analyses offered by these papers. What drives their results is scientists' expectation that genuine scientific achievement will be recognized with credit. As we have argued above, it is long-run credit that best tracks genuine scientific achievement, and so it is long-run rather than short-run credit that grounds scientists' expectation in this regard. So in social epistemologists' most prominent defenses of the credit economy of science, long-run credit (while not named such) is the mechanism underlying the claimed epistemic benefits of the credit economy.

5 Difficulties For Our Proposal

We have discussed some benefits that would predictably accrue from abolishing peer review and some ways in which its apparent benefits are either under-evidenced or better attributed to the effects of long-run credit, which our proposal leaves untouched. We now discuss some cases which we take to be more problematic for our proposal—but by this point we hope to have at least convinced the reader that pre-publication peer review rests on shakier theoretical grounds than its widespread acceptance may lead one to suppose.

5.1 A Guarantee For Outsiders

One purpose pre-publication peer review serves is providing a guarantee to interested but non-expert parties. Science journalists, policy makers, scientists from outside the field the manuscript is aimed at, or interested non-scientists can take the fact that something has passed peer review as a stamp of approval from the field. At a minimum, peer review guarantees that outsiders are focusing on work that has convinced at least one relatively disinterested expert that the manuscript is worthy of public viewing. Given that there are real dangers to irresponsible science journalism or public action that is seen to be based on science that is not itself trustworthy (Bright 2018, §4), and that it is hard for non-experts to make the relevant judgment calls themselves, having a social mechanism to provide this kind of guarantee for outsiders is useful.

It is difficult to predict in advance what norms would come to exist for science journalists in the absence of pre-publication peer review. We thus first and foremost call for empirical research on this issue, possibly by studying what has happened in parts of mathematics and physics that already operate broadly along the lines we suggest (Gowers 2017).

However, against the presumption that things would be worse, we have two points to make. As the recent replication crisis has made clear, the value of peer review as a stamp of approval should not be overstated. There are reasons to doubt that peer review reliably succeeds in filtering out false results. We give three of them. First, peer reviewers face difficulties in actually assessing manuscripts—and just about anything can pass peer review eventually—as discussed under the heading of 'epistemic sorting' in section 4.1. Second, there are problems with the standards we presently use to evaluate manuscripts, in particular with the infamous threshold for statistical significance used in many fields (Ioannidis 2005, Benjamin et al. 2018). And third, deeper features of the incentive structure of science make replicability problems endemic (Smaldino and McElreath 2016, Heesen 2017c). Using

peer review as a stamp of approval may just be generating expert overconfidence (Angner 2006), without the epistemic benefits of greater reliability that would back this confidence up.

For the second part of our reply, recall that it is only pre-publication peer review that we seek to eliminate. We do not object to post-publication peer review resulting in papers being selected for inclusion in journals which mark the community's approval of such work, ideally after due and broad-based evaluation. If some such system were implemented then outsiders could use inclusion in such a journal as their marker of whether work is soundly grounded in the relevant science.

If such a stamp of approval from a journal or other communally recognized institution only comes a number of months or years after something is first published then we would expect it to represent a more well-considered judgment. Note that this would not necessarily slow the diffusion of knowledge as under the present system the same paper would have spent time hidden from view going through pre-publication peer review. The end result might not even be all that different from what happens in the present system, except that post-publication peer review would take into account more of the response or uptake from the wider scientific community. Thus it would more closely approximate the considered judgment of the community, as ultimately reflected in the long-run credit accorded to the paper.

5.2 A Runaway Matthew Effect

The second problem we are less confident we can deal with is that of exacerbating the Matthew effect. This is the phenomenon, first identified by Merton (1968), of antecedently more famous authors being credited more for work done simultaneously or collaboratively, even if the relative size or skill of their contribution does not warrant a larger share of the reward. Arguably the present system helps put a damper on the Matthew effect, allowing a junior or less prestigious author to secure attention for their work by publishing

in a high profile journal. Without such a mechanism to grab the attention of the field, perhaps scientists would just decide what to pay attention to based on their prior knowledge of the author or recommendation from others. This would strengthen the effects of networks of patronage and prestige bias favoring fancy universities. Thus squandering valuable opportunities to learn from those who were not initially lucky in securing a prestigious position or patronage from the already established.

While some have defended the Matthew effect (Strevens 2006), we will not go that route in defending our proposal for two reasons. First, the Matthew effect can perpetuate iniquities that themselves harm the generation and dissemination of knowledge (Bruner and O'Connor 2017). Second, even if it could be justified at the level of individual publications, its long-term effects are epistemically harmful. The scientific community allocates the resources necessary for future work on the basis of its recognition of past performance. So if there is excess reward for some and unfair passing over of others at the present stage of inquiry, this will ramify through to future rounds of inquiry, misallocating resources to people whose accomplishments do not fully justify their renown (Heesen 2017a). Hence on grounds of epistemic consequentialism we take seriously the problem of a runaway Matthew effect.

As mentioned, due to the pressures of credit-seeking and their own curiosity, scientists would still have incentive to read others' work and adapt it to suit their own projects. There is always a chance that valuable knowledge may be gathered from the work of one who has been ignored, which could provide an innovative edge. To some extent this creates opportunities for arbitrage: if the Matthew effect ever became especially severe there would be a credit incentive to specialize in seeking out the work of scientists who are not getting much attention. The lesson here is that the Matthew effect can only ever be so severe, before the credit incentive starts providing counter-veiling motivations.

However, this does not fully solve our problem. Moreover, so long as

resource allocation is tied to recognition of past performance the differences in recognition generated by the Matthew effect can and often do become self-fulfilling prophesies, as those with more gain the resources to do better in the future, and those without are starved of the resources necessary to show their worth.

It is not clear where to go from here. From the above it may seem like a solution would be to pair our proposal with a call to loosen the connection between recognition of a scientist's greatness based on their past performance and resource allocation. Indeed, this may well be independently motivated (Avin forthcoming, Heesen 2017a, §6). However, even short of this far-reaching change, we feel at present that this matter deserves more study rather than any definitive course of action.

Our present thought is that this is a very speculative objection, and there is no empirical evidence to back up the claim that eliminating pre-publication peer review will have dire consequences in this regard. In particular, while the present system may (rarely) allow a relative outsider to make a big splash, the common accusation of prestige bias in peer review (Lee et al. 2013, 7) suggests that on the whole pre-publication peer review may contribute to the Matthew effect rather than curtailing it.

More specifically, the Matthew effect can be made worse by peer review when anonymity breaks down in ways that systematically favor antecedently famous scientists. If this gives famous scientists more opportunities to publish papers, then our system may provide welcome relief, since it allows more people to get their papers out there. Hence whether our proposal makes the Matthew effect worse or better depends on whether the stronger influence would be who gets into the conversation (for which pre-publication peer review can exacerbate the Matthew effect), or who gets listened to once the conversation has begun (for which our proposal looks more problematic). Presently we cannot say which is the more significant effect. So, while we grant that a runaway Matthew effect may occur under our system, we prefer

to stress that at this point it is just not known whether the Matthew effect will be worse with or without pre-publication peer review.

What we propose is a large change, involving freeing up a lot of time and opening it up to more self-direction on the part of scientists, and it is not clear what sort of institutional changes it would be paired with. With more study of epistemic mechanisms designed especially to promote the work of junior or less prestigious scientists there might be found some way of surmounting the problem of a runaway Matthew effect, should it arise. Ultimately, only empirical evidence can settle these questions. Given the clear benefits and the unclear downsides of our proposal, we hope at minimum to inspire a more experimental attitude towards peer review.

6 Conclusion

Pre-publication peer review is an enormous sink of scientists' time, effort, and resources. Adopting the perspective of epistemic consequentialism and reviewing the literature on the philosophy, sociology, and social epistemology of science, we have argued that we can be confident that there would be benefits from eliminating this system, but have no strong reasons to think there will be disadvantages. There is hence a kind of weak dominance or Pareto argument in favor of our proposal.

To simplify things, imagine forming a decision matrix, with rows corresponding to 'Keeping pre-publication peer review' and 'Eliminating pre-publication peer review'. The columns would each be labeled with an issue studied by science scholars which we have surveyed here: gender bias in the literature, speed of dissemination of knowledge, efficient allocation of scientists' time and attention, etc. For each column, if there is a clear reason to think that either keeping or eliminating pre-publication scientific peer review does better according to the standards of epistemic consequentialism, place a 1 in the row of that option, and a 0 in the other. If there is no reason to

favor either according to present evidence, put a 0 in both rows.

Our present argument could then be summarized with: as it stands, the only 1s in such a table would appear in the row for eliminating pre-publication peer review. We thus advocate eliminating pre-publication peer review. Journals could still exist as a forum for recognizing and promoting work that the community as a whole perceives as especially meritorious and wishes to recommend to outsiders. Scientists would still have every reason to read, respond to, and consider the work of their peers; pre-publication peer review is not the primary drive behind either the intellect's curiosity or the will's desire for recognition, and either of those suffice to motivate such behaviors.

The overall moral to be drawn mirrors that of our invocation of the importance of long-run over short-run credit. The best guarantor of the long run epistemic health of science is science: the organic engagement with each others' ideas and work that arises from scientists deciding for themselves how to allocate their cognitive labor, and doing the hard work of replicating and considering from new angles those ideas that have been opened up to the scrutiny of the community. All this would continue without pre-publication peer review, and the best you can say for the system that currently uses up so much of our time and resources is that it often fails to get in the way.

References

Eric Abrahamson. Necessary conditions for the study of fads and fashions in science. *Scandinavian Journal of Management*, 25(2):235–239, 2009. doi: 10.1016/j.scaman.2009.03.005. URL `http://dx.doi.org/10.1016/j.scaman.2009.03.005`.

Jason McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. Epistemic landscapes, optimal search, and the division of cognitive

labor. *Philosophy of Science*, 82(3):424–453, 2015. doi: 10.1086/681766. URL http://dx.doi.org/10.1086/681766.

Melissa S. Anderson, Emily A. Ronning, Raymond De Vries, and Brian C. Martinson. Extending the Mertonian norms: Scientists' subscription to norms of research. *The Journal of Higher Education*, 81(3):366–393, 2010. ISSN 1538-4640. doi: 10.1353/jhe.0.0095. URL https://muse.jhu.edu/journals/journal_of_higher_education/v081/81.3.anderson.html.

Melissa S. Anderson, Marta A. Shaw, Nicholas H. Steneck, Erin Konkle, and Takehito Kamata. Research integrity and misconduct in the academic profession. In Michael B. Paulsen, editor, *Higher Education: Handbook of Theory and Research*, volume 28, chapter 5, pages 217–261. Springer, Dordrecht, 2013. doi: 10.1007/978-94-007-5836-0_5. URL http://dx.doi.org/10.1007/978-94-007-5836-0_5.

Erik Angner. Economists as experts: Overconfidence in theory and practice. *Journal of Economic Methodology*, 13(1):1–24, 2006. doi: 10.1080/13501780600566271. URL http://dx.doi.org/10.1080/13501780600566271.

Shahar Avin. Centralised funding and epistemic exploration. *The British Journal for the Philosophy of Science*, forthcoming. doi: 10.1093/bjps/axx059. URL http://dx.doi.org/10.1093/bjps/axx059.

Nachman Ben-Yehuda. Deviance in science: Towards the criminology of science. *British Journal of Criminology*, 26(1):1–27, 1986. doi: 10.1093/oxfordjournals.bjc.a047577. URL http://dx.doi.org/10.1093/oxfordjournals.bjc.a047577.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical

significance. *Nature Human Behaviour*, 2(1):6–10, 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0189-z. URL http://dx.doi.org/10.1038/s41562-017-0189-z.

Lutz Bornmann. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245, 2011. ISSN 1550-8382. doi: 10.1002/aris.2011.1440450112. URL http://dx.doi.org/10.1002/aris.2011.1440450112.

Thomas Boyer. Is a bird in the hand worth two in the bush? Or, whether scientists should publish intermediate results. *Synthese*, 191(1): 17–35, 2014. ISSN 0039-7857. doi: 10.1007/s11229-012-0242-4. URL http://dx.doi.org/10.1007/s11229-012-0242-4.

Thomas Boyer-Kassem and Cyrille Imbert. Scientific collaboration: Do two heads need to be more than twice better than one? *Philosophy of Science*, 82(4):667–688, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/682940.

John M. Braxton. Deviance from the norms of science: A test of control theory. *Research in Higher Education*, 31(5):461–476, 1990. doi: 10.1007/BF00992713. URL http://dx.doi.org/10.1007/BF00992713.

Liam Kofi Bright. On fraud. *Philosophical Studies*, 174(2):291–310, 2017a. ISSN 1573-0883. doi: 10.1007/s11098-016-0682-7. URL http://dx.doi.org/10.1007/s11098-016-0682-7.

Liam Kofi Bright. Decision theoretic model of the productivity gap. *Erkenntnis*, 82(2):421–442, 2017b. ISSN 1572-8420. doi: 10.1007/s10670-016-9826-6. URL http://dx.doi.org/10.1007/s10670-016-9826-6.

Liam Kofi Bright. Du Bois' democratic defence of the value free ideal. *Synthese*, 195(5):2227–2245, 2018. ISSN 1573-0964.

doi: 10.1007/s11229-017-1333-z. URL http://dx.doi.org/10.1007/s11229-017-1333-z.

Liam Kofi Bright, Haixin Dang, and Remco Heesen. A role for judgment aggregation in coauthoring scientific papers. *Erkenntnis*, 83(2):231–252, 2018. ISSN 1572-8420. doi: 10.1007/s10670-017-9887-1. URL http://dx.doi.org/10.1007/s10670-017-9887-1.

Justin Bruner and Cailin O'Connor. Power, bargaining, and collaboration. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 7, pages 135–157. Oxford University Press, Oxford, 2017.

Justin P. Bruner. Policing epistemic communities. *Episteme*, 10(4):403–416, Dec 2013. ISSN 1750-0117. doi: 10.1017/epi.2013.34. URL http://dx.doi.org/10.1017/epi.2013.34.

Erwin Chargaff. Triviality in science: A brief meditation on fashions. *Perspectives in Biology and Medicine*, 19(3):324–333, 1976. doi: 10.1353/pbm.1976.0011. URL http://dx.doi.org/10.1353/pbm.1976.0011.

Diana Crane. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 2(4):195–201, 1967. ISSN 00031232. URL http://www.jstor.org/stable/27701277.

Partha Dasgupta and Paul A. David. Toward a new economics of science. *Research Policy*, 23(5):487–521, 1994. ISSN 0048-7333. doi: 10.1016/0048-7333(94)01002-1. URL http://www.sciencedirect.com/science/article/pii/0048733394010021.

Margaret Eisenhart. The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, 32(2):241–255, 2002. ISSN 1573-1898. doi: 10.1023/A:1016082229411. URL http://dx.doi.org/10.1023/A:1016082229411.

Edzard Ernst, T. Saradeth, and Karl Ludwig Resch. Drawbacks of peer review. *Nature*, 363(6427):296, 1993. doi: 10.1038/363296a0. URL `http://dx.doi.org/10.1038/363296a0`.

Henry Etzkowitz, Stefan Fuchs, Namrata Gupta, Carol Kemelgor, and Marina Ranga. The coming gender revolution in science. In Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, editors, *The Handbook of Science and Technology Studies*, chapter 17, pages 403–428. MIT Press, Cambridge, third edition, 2008. ISBN 9780262083645.

Daniele Fanelli. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, 5(4):e10271, Apr 2010. doi: 10.1371/journal.pone.0010271. URL `http://dx.doi.org/10.1371/journal.pone.0010271`.

Jere R. Francis. The credibility and legitimation of science: A loss of faith in the scientific narrative. *Accountability in Research: Policies and Quality Assurance*, 1(1):5–22, 1989. doi: 10.1080/08989628908573770. URL `http://dx.doi.org/10.1080/08989628908573770`.

Alvin I. Goldman. *Knowledge in a Social World*. Oxford University Press, Oxford, 1999. ISBN 0198237774.

Timothy Gowers. The end of an error? *The Times Literary Supplement*, October 2017. URL `https://www.the-tls.co.uk/articles/public/the-end-of-an-error-peer-review/`. Editorial.

Paul M. Grant. Scientific credit and credibility. *Nature Materials*, 1:139–141, 2002. doi: 10.1038/nmat756. URL `http://dx.doi.org/10.1038/nmat756`.

Sandra Harding. "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3):331–349, 1995. doi: 10.1007/BF01064504. URL `http://dx.doi.org/10.1007/BF01064504`.

Remco Heesen. Academic superstars: Competent or lucky? *Synthese*, 194 (11):4499–4518, 2017a. ISSN 1573-0964. doi: 10.1007/s11229-016-1146-5. URL http://dx.doi.org/10.1007/s11229-016-1146-5.

Remco Heesen. Communism and the incentive to share in science. *Philosophy of Science*, 84(4):698–716, 2017b. ISSN 0031-8248. doi: 10.1086/693875. URL http://dx.doi.org/10.1086/693875.

Remco Heesen. Why the reward structure of science makes reproducibility problems inevitable. Manuscript, September 2017c. URL http://remcoheesen.files.wordpress.com/2015/03/rewards-and-reproducibility2.pdf.

Remco Heesen. When journal editors play favorites. *Philosophical Studies*, 175(4):831–858, 2018. ISSN 0031-8116. doi: 10.1007/s11098-017-0895-4. URL http://dx.doi.org/10.1007/s11098-017-0895-4.

Remco Heesen, Liam Kofi Bright, and Andrew Zucker. Vindicating methodological triangulation. *Synthese*, forthcoming. ISSN 1573-0964. doi: 10.1007/s11229-016-1294-7. URL http://dx.doi.org/10.1007/s11229-016-1294-7.

Erin Hengel. Publishing while female: Are women held to higher standards? Evidence from peer review. Manuscript, August 2018. URL http://www.erinhengel.com/research/publishing_female.pdf.

David L. Hull. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. The University of Chicago Press, Chicago, 1988. ISBN 0226360504.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, Aug 2005. doi: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

Saana Jukola. A social epistemological inquiry into biases in journal peer review. *Perspectives on Science*, 25(1):124–148, 2017. doi: 10.1162/POSC_a_00237. URL http://dx.doi.org/10.1162/POSC_a_00237.

J. Katzav and K. Vaesen. Pluralism and peer review in philosophy. *Philosophers' Imprint*, 17(19):1–20, 2017. URL http://hdl.handle.net/2027/spo.3521354.0017.019.

Philip Kitcher. The division of cognitive labor. *The Journal of Philosophy*, 87(1):5–22, 1990. ISSN 0022362X. URL http://www.jstor.org/stable/2026796.

Richard L. Kravitz, Peter Franks, Mitchell D. Feldman, Martha Gerrity, Cindy Byrne, and William M. Tierney. Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? *PLoS ONE*, 5(4):e10072, 2010. doi: 10.1371/journal.pone.0010072. URL http://dx.doi.org/10.1371/journal.pone.0010072.

Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, second edition, 1986.

Carole J. Lee. The limited effectiveness of prestige as an intervention on the health of medical journal publications. *Episteme*, 10(4):387–402, 2013. doi: 10.1017/epi.2013.35. URL http://dx.doi.org/10.1017/epi.2013.35.

Carole J. Lee. Revisiting current causes of women's underrepresentation in science. In Jennifer Saul and Michael Brownstein, editors, *Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology*, chapter 2.5, pages 265–282. Oxford University Press, Oxford, 2016. doi: 10.1093/acprof:oso/9780198713241.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780198713241.001.0001.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

Christin List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001. ISSN 1467-9760. doi: 10.1111/1467-9760.00128. URL http://dx.doi.org/10.1111/1467-9760.00128.

Helen E. Longino. *Science as Social Knowledge*. Princeton University Press, 1990.

Karen Seashore Louis, Lisa M. Jones, and Eric G. Campbell. Macroscope: Sharing in science. *American Scientist*, 90(4):304–307, 2002. ISSN 00030996. URL http://www.jstor.org/stable/27857685.

Bruce Macfarlane and Ming Cheng. Communism, universalism and disinterestedness: Re-examining contemporary support among academics for Merton's scientific norms. *Journal of Academic Ethics*, 6(1):67–78, 2008. ISSN 1570-1727. doi: 10.1007/s10805-008-9055-y. URL http://dx.doi.org/10.1007/s10805-008-9055-y.

Robert K. Merton. A note on science and democracy. *Journal of Legal and Political Sociology*, 1(1–2):115–126, 1942. Reprinted in Merton (1973, chapter 13).

Robert K. Merton. Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22(6):635–659, 1957. ISSN 00031224. URL http://www.jstor.org/stable/2089193. Reprinted in Merton (1973, chapter 14).

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63,

1968. ISSN 00368075. URL http://www.jstor.org/stable/1723414. Reprinted in Merton (1973, chapter 20).

Robert K. Merton. Behavior patterns of scientists. *The American Scholar*, 38 (2):197–225, 1969. ISSN 00030937. URL http://www.jstor.org/stable/41209646. Reprinted in Merton (1973, chapter 15).

Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago, 1973. ISBN 0226520919.

Brian A. Nosek, Jeffrey R. Spies, and Matt Motyl. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012. doi: 10.1177/1745691612459058. URL http://pps.sagepub.com/cgi/content/abstract/7/6/615.

Cailin O'Connor and Justin Bruner. Dynamics and diversity in epistemic communities. *Erkenntnis*, forthcoming. ISSN 1572-8420. doi: 10.1007/s10670-017-9950-y. URL http://dx.doi.org/10.1007/s10670-017-9950-y.

Slobodan Perović, Sandro Radovanović, Vlasta Sikimić, and Andrea Berber. Optimal research team composition: data envelopment analysis of Fermilab experiments. *Scientometrics*, 108(1):83–111, 2016. doi: 10.1007/s11192-016-1947-9. URL http://dx.doi.org/10.1007/s11192-016-1947-9.

Douglas P. Peters and Stephen J. Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2):187–195, 1982. doi: 10.1017/S0140525X00011213. URL http://dx.doi.org/10.1017/S0140525X00011213.

Katarina Prpić. Gender and productivity differentials in science. *Scientometrics*, 55(1):27–58, 2002. ISSN 0138-9130. doi: 10.1023/A:1016046819457. URL http://dx.doi.org/10.1023/A:1016046819457.

RIN. Activities, costs and funding flows in the scholarly communications system in the UK. Technical report, Cambridge Economic Policy Associates on behalf of the Research Information Network, 2008. URL http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/activities-costs-and-funding-flows-scholarly-commu.

Felipe Romero. Novelty versus replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, 84(5):1031–1043, 2017. ISSN 0031-8248. doi: 10.1086/694005. URL http://dx.doi.org/10.1086/694005.

Jennifer Saul. Implicit bias, stereotype threat, and women in philosophy. In Katrina Hutchison and Fiona Jenkins, editors, *Women in Philosophy: What Needs to Change?*, chapter 2, pages 39–60. Oxford University Press, Oxford, 2013.

Paul E. Smaldino and Richard McElreath. The natural selection of bad science. *Royal Society Open Science*, 3(9), 2016. doi: 10.1098/rsos.160384. URL http://rsos.royalsocietypublishing.org/content/3/9/160384.

Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182, 2006. URL http://jrs.sagepub.com/content/99/4/178.short.

Paula E. Stephan. The economics of science. *Journal of Economic Literature*, 34(3):1199–1235, 1996. URL http://www.jstor.org/stable/2729500.

Michael Strevens. The role of the priority rule in science. *The Journal of Philosophy*, 100(2):55–79, 2003. ISSN 0022362X. URL http://www.jstor.org/stable/3655792.

Michael Strevens. The role of the Matthew effect in science. *Studies in History and Philosophy of Science Part A*, 37(2):159–170, 2006. ISSN 0039-3681. doi: http://dx.doi.org/10.1016/j.shpsa.2005.07.009. URL http://www.sciencedirect.com/science/article/pii/S0039368106000252.

Michael Strevens. Herding and the quest for credit. *Journal of Economic Methodology*, 20(1):19–34, 2013. doi: 10.1080/1350178X.2013.774849. URL http://dx.doi.org/10.1080/1350178X.2013.774849.

Michael Strevens. Scientific sharing: Communism and the social contract. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 1. Oxford University Press, Oxford, 2017. URL https://philpapers.org/rec/STRSSC-2.

Johanna Thoma. The epistemic division of labor revisited. *Philosophy of Science*, 82(3):454–472, 2015. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/681768.

Virginia Valian. *Why So Slow? The Advancement of Women*. MIT Press, Cambridge, 1999. ISBN 9780262720311.

Richard Van Noorden. The true cost of science publishing. *Nature*, 495 (7442):426–429, 2013. ISSN 0028-0836. doi: 10.1038/495426a. URL http://dx.doi.org/10.1038/495426a.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/644786.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6(2):185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://dx.doi.org/10.3366/E174236000900063X.

Kevin J. S. Zollman. The credit economy and the economic rationality of science. *The Journal of Philosophy*, 115(1):5–33, 2018. doi: 10.5840/jphil201811511. URL http://dx.doi.org/10.5840/jphil201811511.

Harriet Zuckerman and Jonathan R. Cole. Women in American science. *Minerva*, 13(1):82–102, 1975. ISSN 1573-1871. doi: 10.1007/BF01096243. URL http://dx.doi.org/10.1007/BF01096243.

Epistemic Loops and Measurement Realism

Alistair M. C. Isaac

Abstract

Recent philosophy of measurement has emphasized the existence of both diachronic and synchronic "loops," or feedback processes, in the epistemic achievements of measurement. A widespread response has been to conclude that measurement outcomes do not convey interest-independent facts about the world, and that only a coherentist epistemology of measurement is viable. In contrast, I argue that a form of measurement realism is consistent with these results. The insight is that antecedent structure in measuring spaces constrains our empirical procedures such that successful measurement conveys a limited, but veridical knowledge of "fixed points," or stable, interest-independent features of the world.

§1 Introduction

Recent philosophy of measurement has employed detailed case studies to highlight the complex, iterative process by which measurement practices are refined. Typically, these examples are taken to support some form of epistemic coherentism, on which the validation of measurement procedures, and thus their epistemic import, is irreducibly infected by the contingent history of their development in aid of human interests. This coherentism in turn undermines *measurement realism*, the view that outcomes of successful measurement practices veridically represent objective (i.e. interest-independent) features of the world. For instance, van Fraassen (2008) takes the historical contingency of measurement practice to support empiricism, and Chang (2012) argues that only a pragmatic, interest-relative "realism" about measurement outcomes is plausible, not one which interprets them as corresponding to objective features in the world. More generally, Tal (2013) identifies coherentism as a major trend within contemporary philosophy of measurement.

I argue that the iterative and coherentist features of measurement practice these authors rightly emphasize are nevertheless consistent with realism about measurement outcomes. Nevertheless, my position contrasts significantly with that of other measurement realists, such as Byerly and Lazara (1973) or Michell (2005), who take measurement realism to be continuous with global scientific realism. On their view, measurement realism is a *stronger* position than traditional realism, imputing reality not only to theoretical objects and laws, but also to their quantitative character. The view defended here reverses this priority, articulating a realism about measurement outcomes *weaker* than traditional realism. In particular, I argue that the convergent assignment of increasingly precise values that constitutes successful measurement serves as incontrovertible evidence for *fixed points* in the world — features or events standing in stable quantitative relationships — even though the evidence it provides for any non-numerical theoretical description of these points is defeasible. The insight here is that measurement is more evidentially demanding than traditional confirmation, i.e. it requires a greater contribution from the interest-independent world to succeed than mere qualitative experiments. I argue that this greater evidential demand is a consequence of the

antecedent numerical structure in which measurement outcomes are represented. This antecedent structure blocks the possibility of gerrymandered categories that crosscut the joints of nature. Consequently, successful measurement constitutes a substantive enough epistemic achievement that we may legitimately "factor out" the contribution to success made by human interests, and accept the outcome as representing an objective feature of the world.

After surveying the motivations for measurement coherentism, I elaborate on the notion of "successful" measurement, and why it poses a challenge to coherentism. The paper concludes with a more careful articulation of the distinctive features of fixed point realism.

§2 Epistemic Loops in Measurement Practice

Contemporary measurement coherentism is motivated by two types of case study, each identifying a different kind of epistemic "loop," or feedback process driving knowledge formation. Chang and van Fraassen emphasize diachronic examples of epistemic iteration, where the feedback process extends over several stages of mutual influence between theory change and refinement of measurement practice. A different kind of epistemic loop has been discussed by Tal and metrologist Mari, who highlight the role of models in the calibration of measurement instruments and the assignment of quantity values, illustrating a synchronic epistemic interdependence between theory and measurement.

§2.1 Epistemic Iteration

Chang (2004) defines *epistemic iteration* as "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals" (45). He takes this process to support a "progressive coherentism": on the one hand, the criteria for measurement success are internal to a practice, so scientific knowledge does not rest on an independent foundation; on the other hand, these internal criteria may be used to evaluate new practices as improvements or refinements on their predecessors, thereby allowing for scientific progress (in contrast to traditional coherentism, Chang 2007). In the context of measurement, this means that later measurement practices may be understood as in some sense "better" than earlier ones, yet these "epistemic achievements" should not be cashed out as greater degree of correspondence to quantities in the world.

For instance, thermometry as a practice begins with subjective assignments of relative heat on the basis of our bodily experiences. Noticing that fluids appear to change volume in rough correspondence with these subjective sensations, one may construct a thermoscope, or device allowing comparison of relative fluid volumes in different circumstances. Already a theoretical leap is required to identify the cause of these changes in relative volume with the cause of our differing subjective sensations, especially given the discrepancies between these sensations and our thermoscopic readings (e.g. contrary to experience, caves are warmer in summer than they are in winter). Nevertheless, the move to the thermoscope constitutes an epistemic achievement, in the sense that it allows for greater regularity in the assignment of relative temperatures, both

across contexts and across observers. A similar pattern is seen in the move from thermoscope to thermometer, which enables assignment of numbers to temperatures. Numerical representation constitutes a yet greater epistemic achievement, insofar as it allows comparison of temperature assignments across devices. Nevertheless, this practice does not itself guarantee greater veracity of temperature assignments, since it rests on the assumption that temperature varies linearly with changes in the height of thermometric fluid. But this assumption cannot itself be verified, as that would require access to temperature in the world by some means independent of thermometry. Similar achievements, (seemingly) inextricably entangled with theory, may be seen at each further stage in the development of thermometric practice.

The moral of this case study is the historical contingency of thermometry, and thus of its results. At each stage in the development of thermometry, an advance in theory was required to extend measurement practice. Internal criteria of consistency and increased precision in the assignment of numerical values establish the new practice as an advance over the previous one. Yet, the application of these criteria is not empirically constrained. When one assumes that "temperature" (whatever it may be) varies linearly with changes in the height of the indicator column in an air thermometer, one is making an assumption both necessary for measurement progress and in principle non-empirical, since no independent access to "temperature," outside the behavior of the very devices and procedures under investigation, is possible: "*Prior to the construction of a thermometer, there is no thermometer to settle that question!*" (van Fraassen 2008, 126, emphasis in original). Chang (2004) argues that, in order to make sense of the "progress" exemplified by cases like these, we have to "look away from truth," and appeal only to historically contingent criteria for success (227)—"scientific progress … cannot mean closer approach to the truth" (228); "Truth, in the sense of correspondence to reality, is beyond our reach" (Chang 2007, 20). The delusion that one may evaluate the correspondence between our assignment of temperatures and the objective state of the world rests on the mistaken and "impossible god-like view in which nature and theory and measurement practice are all accessed independently of each other" (van Fraassen 2008, 139). Rather, the only relevant notion of "truth" for assessing the success of thermometry "rests first and foremost on coherence with the rest of the system" (Chang 2012, 242).

§2.2 Models and Calibration

Another kind of epistemic loop is found in synchronic measurement practice, where *models* play a constitutive role in determining measurement outcomes. The crucial concept here is *calibration*, the process of correcting a measurement device for inferred discrepancies between its readout and the target value. Calibration is a necessary feature of all sophisticated measurement, yet the process of calibration illustrates the ineliminable role of theoretical posits in the very assignment of quantity values in an act of measurement. When measuring, scientists do not (as one might naively suppose) read values directly from nature, rather they employ models of the interaction between measurement device and target system in order to "correct" the readout value to a final assigned value (Mari and Giordani 2014).

Tal (2014) illustrates this point through the example of the measurement of time, in particular coordinated universal time (UTC). The second is presently defined as 9,192,631,770 periods of the hyperfine transition between the two ground states of a caesium-133 atom at zero degrees Kelvin.[1] Models feature at every step of the process leading from devices that interact directly with caesium atoms to the UTC. First, it is impossible to probe caesium atoms at absolute zero, so the enumeration of hyperfine transitions output by a caesium clock must be corrected for this discrepancy. This, as well as other corrections, rely on models of the physical interaction between the device and the atom in order to infer the discrepancy between the actual state of the system and the idealized state referred to in the definition. Caesium clocks are too complex to run continuously, so their output is used to calibrate more mundane atomic clocks (301). Furthermore, the UTC itself is not identified with the output of any one clock; rather, it is calculated retrospectively by a weighted average over all participating atomic clocks, with weights determined by the degree of past fit between each clock and previous calculations of UTC (302–3).

The lessons of this example are analogous to those of epistemic iteration: measurement improvement appears to rest on internal standards of coherence rather than on correspondence with external quantities. The weighting procedure that leads to UTC, for instance, "promotes clocks that are stable relative to each other" (304). Success at achieving this stability indeed demonstrates "genuine empirical knowledge," but not knowledge in the first instance about a regularity in the objective world, but rather a regularity "in the behaviour of instruments" (327). Consequently, it is a "conceptual mistake" to think that "the stability of measurement standards can be analysed into distinct contributions by humans and nature" (328). On an extreme interpretation of this view, even computer simulation constitutes a form of measurement (Morrison 2009). The basic idea is that, once we grant the ineliminable role of models in measurement, it is a small conceptual step to accept that the aspect of measurement involving empirical contact with the world may be arbitrarily distant from that involving modeling (Parker 2017).

§3 Achieving Successful Measurement

For the remainder of this paper, I wish to grant the basic descriptive features of this account: both diachronically and synchronically, successful measurement involves epistemic loops. Nevertheless, I will argue, there is a form of measurement realism consistent with these loops; one on which the contingent, interest-relative, and theory-laden aspects of measurement may indeed be factored out, leaving the bare, objective facts about the world conveyed by successful measurement.

[1] Arguably, the process of establishing UTC is not measurement at all — since the length of the second is *defined* by caesium-133 transitions, it is not subject to empirical determination. The purpose of the project Tal examines is not to establish a value, as in paradigmatic cases of measurement, but rather to coordinate time-relevant activities across the globe with maximal precision. I set this concern aside for the discussion here, since Tal's analysis has been so influential in philosophy of measurement, and his conclusions concerning the model-mediation of measurement incontrovertibly reflect the practices of metrologists.

But what is "successful measurement"? For the purposes of discussion here, I take *measurement* to be any empirical procedure for assigning points (or regions) in a metric space to states of the world, where a *metric space* is any set of elements with a distance metric defined over it. This means, on the one hand, that I rule out degenerative forms of "measurement" that simply assign objects to categories, or place them in an order (the *nominal* and *ordinal* scales of Stevens 1946). On the other hand, I include measurement procedures that map states of the world into any geometrical space, not just the real line, so long as they have an assigned distance metric (siding with Suppes, et al. 1989, against Díez 1997); nevertheless, in the interests of simplicity, I will refer to these outcomes as "numerical" assignments, since they may be represented by vectors of real numbers. In line with Krantz et al. (1971), I take it that one can determine whether or not an empirical procedure constitutively requires the metric features of a geometrical space by analyzing whether these remain invariant across permissible transformations over the mapping into that space.[2]

I take *successful* measurement to exhibit two key features: *convergence* and *precision*. These features pose a significant challenge to the thoroughgoing coherentist.

§3.1 Convergence

Coherentists have emphasized the theory-ladenness of both diachronic and synchronic aspects of measurement refinement. However, a hallmark of sophisticated scientific measurement is its attempt to factor out the role of theory in measurement by employing different theoretical commitments to measure the same quantity. A measurement practice *converges* when procedures employing different theoretical commitments arrive at the same outcome.

For instance, in the early 20th century, a wide variety of phenomena were investigated, employing distinct methods and theoretical commitments, in the attempt to measure Avogadro's constant N_A, the number of particles in a mole of substance. Perhaps most well-known are Perrin's experiments on Brownian motion, which, in combination with Einstein's theoretical analysis, allowed an assignment of value to N_A. However, similar values were achieved by radically different means. For instance, Millikan was able to determine N_A by measuring charge of the electron through his oil drop experiments and dividing the Faraday constant (charge of a mole of electrons) by his result. Millikan's measurement relied on Stokes' theoretical analysis of the movement of spheres through a viscous fluid — insofar as Brownian motion was a factor, it was as a source of noise, not (as for Perrin) a source of evidence. Black body radiation and the blue

[2] For instance, consider two procedures for assigning real numbers to my students. On the first, I assign a number to each letter-type with which a student's name begins (e.g. A=1, B=3,…); on the second, I hold a meter stick up to each student and note their height. The former procedure is indifferent to the algebraic structure of the real line (letters do not add or subtract from each other systematically), and thus metric features of the real line are not invariant across alternative, equally permissible assignments of numbers (e.g. A=7, B=15,…). The second does make use of algebraic structure (as heights do "add" through concatenation), and thus metric features remain invariant across alternative assignments (Jamal is twice the height of Leslie, whether their heights are represented in inches or centimeters). So, on the present definition, the latter procedure is measurement, but the former is not.

of the sky are examples of other phenomena that, when combined with theoretical models of photon emission and diffraction respectively, allow alternate means of measuring N_A. Insofar as these procedures assign the same value to N_A, they converge.

I want to stress that the point being made here is *not* the traditional realist one, that these practices provide converging evidence for the particulate nature of matter, whether as "common cause" (Salmon 1984) or most likely hypothesis (Psillos 2011). Those arguments are instances of *abduction*, while I am interested in whether a stronger, non-abductive conclusion may be drawn from convergence. A better analogy is with the discussion of robustness in the modeling literature: a result is *robust* if it is obtained by a plurality of models that each make different simplifying assumptions (Weisberg 2006). The particulate nature of matter is not robust in this sense across different measurement practices, since it is assumed by all of them. However, the value of N_A is robust, since that value is not itself assumed, and is obtained with a great degree of agreement despite differences in the assumptions made by each measurement practice (and its supporting models). I claim that convergence toward this value provides robust, non-abductive evidence for an objective feature of the world.

This example is in no way exceptional: convergent measurement practices are rife across the sciences. Smith and Miyake, for instance, have investigated a number of examples. Thomson's convergent measurements of the charge of the electron employed a variety of different methods and assumptions (Smith 2001). Early attempts to measure the density of the interior of the earth likewise assumed a variety of different theoretical models (Miyake 2018). In more recent research, measurements of the constants that govern molecular vibration converge across spectroscopy, chemistry, thermodynamics, and femtochemistry (Smith and Miyake, *manuscript*). To pick an example from an entirely different area of science, measurements of the spectral sensitivity of mammalian retinal receptors employing psychophysical methods (extracting sensitivity curves from behavioral color matching experiments, as performed by Helmholtz in the late 19th century) converge closely with 20th century physiological methods (detecting rate of nerve firing in (e.g.) cow retinal tissue in response to single wavelength lights, Wandell 1995). In all of these cases, "What is being shown through the convergence of these measurements is that the discrepancies between the different measurements … are due to the particularities of the models being used" (Miyake, 2018, 336). In other words, convergence factors out model-sensitive features of measurement; in order for it to occur, "the empirical world has to cooperate" (Smith 2001, 26).

§3.2 Precision

Traditionally, measurement success was evaluated with respect to two features: accuracy and precision. *Accuracy* was degree of approach to true value, while *precision* was degree of specificity in the value provided. The considerations in §2 undermine the criterion of accuracy, since they show we have no independent access to "true values" and thus cannot use them as standards for evaluating measurement (Mari 2003). Nevertheless, we can still assess measurements for precision, since it may be defined operationally: a measurement is *precise* to the

extent that it returns the same result when performed repeatedly. The number of *significant figures* in a numerical assignment indicates the degree of measurement precision, since these characterize the size of the region within which repeated measurements fall.

Coherentists stress the fact that increased precision is a purely internal criterion for improving measurement. Here, however, I want to stress the way in which increased precision constitutes a qualitatively different, and more impressive, epistemic achievement than other forms of empirical success, such as qualitative prediction or improved coherence of classification. These qualitative achievements are subject to worries about semantic and theoretical holism: one may always succeed in classification, or correct qualitative prediction, by suitably redrawing the boundaries of one's theoretical concepts. As LaPorte (2004) argues, when faced with anomalies in the relationship between guinea pigs and prototypical rodents, or birds and dinosaurs, scientists face a *choice* whether to expand or contract their previous categories to include or exclude perceived outliers (a similar case is made by Slater 2017 for Pluto and planethood). Nothing about the prior conceptual framework itself forces this choice one way or another, nor do demands for internal consistency.

Measurement is different from mere categorization precisely because it maps states into a metric space. The crucial point to note here is that a metric space has *antecedent structure*: the distances between points on the real line, and the algebraic relationships between them, are fixed *before* we employ it to represent height or temperature or electric charge. This antecedent structure constrains the relationship between measurement outcomes, independently restricting our assessment of them as same or different, or converging or not, in a manner impervious to ad hoc revision. Increase in precision occurs when successive measurement practices are able to shrink distances (between repeated measurements within each practice) determined by the metric of the representing space. Thus, the metric of this space serves two functions: (i) it represents the distances between different measured quantities, but (ii) it also provides a directed metric for improving measurement of a single quantity, since it determines the distances between repeated measurements that characterizes their precision. Consequently, pace van Fraassen, attempts to increase precision are empirically constrained, since this directed metric for improvement can only be satisfied through the cooperation of nature: if nature is not sufficiently stable where we probe it, no choice, convention, or increased coherence can reduce the distances between our repeated attempts to measure it. Some examples will illustrate this point.

Consider, for instance, determinations of the boiling point of water. Chang (2004, Ch. 1) surveys the sequence of choice points in the early practice of thermometry leading to relative stability in the measurement of this temperature: what are the visual indicators of boiling, where should the thermometer be positioned, what should be the shape of the vessel holding the water, its material, etc.[3] Decisions on each of these points affect the relative stability in the thermometric reading, illustrating the naivety of a view on which

[3] The issue here is the phenomenon of "superheating," whereby water with relatively little dissolved gas, or in a flask with very small surface area, may be heated to a higher temperature without bubbling.

boiling point is a simple phenomena merely waiting to be observed. Nevertheless, in committing to represent the boiling point numerically, investigators subjected themselves to a criterion for success distinct from coherence. If the numbers assigned by thermometers within this-shaped vessels and that-shaped ones differ during phenomenologically similar bubblings, then the distance between those numbers provides a criterion of difference that must be respected if thermometric practice is to count as measurement. Restricting attention to those vessels that minimize distances between numerical outcomes is thus not a mere choice, or gerrymandering of the category "boiling," since it is forced upon the investigator by an antecedent metric for success.

Likewise, consider again the determination of UTC through the retrospective weighting of the comparison set of atomic clocks. For Tal, the success of this procedure is evidence for stability in our clocks, but not for any human-independent feature of the world. Nevertheless, UTC is constrained by the world in two distinct ways. First, through empirical contact with caesium atoms. While this contact is mediated by models, these models themselves are the result of convergent measurements of atomic phenomena through a wide variety of means, employing distinct theoretical assumptions. Second, the distance metric of the real line constrains the assessment of fit between clocks in the set. While the algorithm that weights them takes degree of internal agreement as the standard for higher weighting, the metrical structure of the space in which relative rates of the clocks are assessed ensures relative agreement cannot be stipulated, fudged, or gerrymandered. The clocks need to cooperate by performing stably enough that they may be compared with a high degree of precision, and this stable point remains tethered to a robust regularity in the world through checks with the convergent behavior of caesium.

While UTC is in some respects atypical (see footnote 1), these three features — internal coordination of outcomes, empirical checks, and directed improvement constrained by the real line — are features of scientific measurement in general. What Tal's discussion of the UTC obscures is the sheer number of empirical checks typically involved, and the strictness of the demands placed by conformity to the metric of improvement the measuring space provides. In official determinations of fundamental physical constants, convergence is demanded across *all* measurement procedures, as assessed by the law-governed interrelationship between physical quantities, and the degree of precision achieved illustrates the strictness of this demand. For instance, in late 19th century measurements of N_A by Perrin and e (charge of electron) by Thomson, only 2 to 3 significant figures were typically obtained within method, and convergence across methods often only agreed as to order of magnitude. By 1911, Millikan was measuring both e and N_A to 4 significant figures, and demonstrating that the models employed to calibrate the oil drop method converged closely with other aspects of physical theory (1911). As of 2014, N_A was being measured at upwards of 9 significant figures, and e upwards of 11 (Mohr et al. 2016).[4] In each case, the increase in precision has been constrained by the antecedent structure of the real line, and thus is not itself a matter of mere convention or coherence. Rather, the world must cooperate by remaining

[4] It is expected that after the 2018 26th General Conference on Weights and Measures, N_A and e will be fixed as constants to which other quantities may be referred during measurement.

sufficiently stable if such precision is to be possible; consequently, precise values constitute robust evidence for points of objective fixity in the world revealed through measurement.

§4 Conclusion: Fixed-Point Realism

Traditional scientific realism rests on an abductive inference from observed empirical success to presumed underlying causes. Successful measurement may certainly be used in such an inference, but I claim here that it non-abductively supports a more modest realism:

> **Fixed Point Realism** – values obtained through successful measurement veridically represent objective fixed points in the world, which may be exhaustively characterized by the pattern of distances that obtain between them in a metric space.

FPR is a form of *epistemic structural realism*. It differs from traditional realism insofar as it claims a veridical characterization of the world is possible independent of any particular theoretical description. Our theory of the nature of temperature or of state changes may change radically, yet the points of relative stability characterizing, e.g., boiling point of water, "absolute zero," freezing point of oxygen, etc., will stay robust across any such change, and that robustness may be represented by their relative positions within a numerical scale.

FPR differs from other flavors of structural realism in the type of structure to which it is committed. Structural realists typically focus on the rich mathematical structure of physical theory, and derivation or limit relations that hold between successive theories, e.g. Newton's laws are a limit case of relativistic mechanics (Worrall 1989). FPR commits itself only to *geometric* structure, i.e. the pattern of relative distances that obtain between points of stability as represented in a metric space. Just as our theoretical description of these stable points may change, so may our mathematical account of their relationship — if new mathematical physics fails to derive old equations as limit cases, this in no way jeopardizes the veridicality of this geometric structure.

Finally, FPR disagrees with coherentism, insofar as it asserts that the geometrical structure uncovered through acts of successive measurement obtains in the world independent of our practices. It does not deny the importance of epistemic loops for understanding the process of measurement. Nevertheless, it takes convergence in measured values to indicate that the points of stability they represent obtain independent of the theoretical commitments encapsulated in the models used for calibration. Likewise, it takes increased precision to constitute a criterion for measurement success over and above that of coherence, one that is only realized when the interest-independent world cooperates with us by remaining stable when we probe it.

Bibliography

Byerly, H., and V. Lazara (1973) "Realist Foundations of Measurement," *Philosophy of Science* 40:10–28.

Chang, H. (2004) *Inventing Temperature*, Oxford UP.

Chang, H. (2007) "Scientific Progress: Beyond Foundationalism and Coherentism," O'Hear (ed.) *Royal Institute of Philosophy Supplement* 61:1–20.

Chang, H. (2012) *Is Water H_2O?* Springer.

Díez, J. (1997) "A Hundred Years of Numbers: An Historical Introduction to Measurement Theory 1887–1990, part ii," *Studies in History and Philosophy of Science* 28:237–265.

Krantz, D., R. Luce, P. Suppes, and A. Tversky (1971) *Foundations of Measurement*, vol. 1, Dover.

LaPorte, J. (2004) *Natural Kinds and Conceptual Change*, Cambridge UP.

Mari, L. (2003) "Epistemology of Measurement," *Measurement* 34:17–30.

Mari, L., and A. Giordani (2014) "Modeling Measurement: Error and Uncertainty," in Boumans, Hon, and Petersen (eds.) *Error and Uncertainty in Scientific Practice*, Pickering & Chatto: 79–96.

Michell, J. (2005) "The Logic of Measurement: A Realist Overview," *Measurement* 38:285–294.

Millikan, R. (1911) "On the Elementary Electrical Charge and the Avogadro Constant," *Physical Review* 2:349–397.

Miyake, T. (2018) "Scientific Realism and the Earth Sciences," in Saatsi (ed.) *The Routledge Handbook of Scientific Realism*, Routledge: 333–344.

Mohr, P., D. Newell, and B. Taylor (2016) "CODATA Recommended Values of the Fundamental Physical Constants: 2014," *Review of Modern Physics* 88:035009.

Morrison, M. (2009) "Models, Measurement and Computer Simulation: The Changing Face of Experimentation," *Philosophical Studies* 143:33–57.

Parker, W. (2017) "Computer Simulation, Measurement, and Data Assimilation," *British Journal for Philosophy of Science* 68:273–304.

Psillos, S. (2011) "Moving Molecules above the Scientific Horizon: On Perrin's Case for Realism," *Journal for General Philosophy of Science* 42:339–363.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton UP.

Slater, M. (2017) "Plato and the Platypus: An Odd Ball and an Odd Duck – On Classificatory Norms," *Studies in History and Philosophy of Science* 61:1–10.

Smith, G. (2001) "J.J. Thomson and the Electron, 1897–1899," in Buchwald and Warwick (eds.) *Histories of the Electron*, MIT Press.

Smith, G., and T. Miyake (*manuscript*) "Realism, Physical Meaningfulness, and Molecular Spectroscopy"

Stevens, S. (1946) "On the Theory of Scales of Measurement," *Science* 103(2684):677–680.

Suppes, P., D. Krantz, R. Luce, and A. Tversky (1989) *Foundations of Measurement*, vol. 2, Dover.

Tal, E. (2013) "Old and New Problems in Philosophy of Measurement," *Philosophy Compass* 8/12:1159–1173.

Tal, E. (2014) "Making Time: A Study in the Epistemology of Measurement," *British Journal for Philosophy of Science* 67:297–335.

Wandell, B. (1995) *Foundations of Vision*, Sinauer.

Weisberg, M. (2006) "Robustness Analysis," *Philosophy of Science* 73:730–742.

Worrall, J. (1989) "Structural Realism: The Best of Both Worlds," *Dialectica* 43:99–124.

van Fraassen, B. (2008) *Scientific Representation*, Oxford UP.

The relationship between intervention and representation is currently resurfacing in philosophy of science. Analytical treatments of the specific intersections between *representation* and *intervention* have recently been explored in Hacking (1983), Radder (2003), Heidelberger (2003), van Fraassen (2008), and Keyser (2017). These accounts analyze intervention-based experimental and measurement practice and the *consequences* for representing and model-building. Of particular interest in my discussion is that some of these accounts explicitly differentiate between representational and productive roles in scientific practice. For example, Heidelberger (2003) and van Fraassen (2008) discuss the representational and productive roles of instruments in experiment and measurement. In the former role, relations in a natural phenomenon are represented in an instrument (van Fraassen 2008, 94). In the latter role, instruments create new phenomena or mimetic phenomena, which resemble natural phenomena. Keyser (2017) takes the distinction between representation and production a step further to differentiate two types of experimental/measurement methodologies:

> When scientists measure/experiment they can *take* measurements, in which case the primary aim is to represent natural phenomena. Scientists can also *make* measurements, in which case the aim is to intervene in order to *produce* experimental objects and processes—characterized as 'effects'. (Keyser 2017, 2)

On Keyser's account 'taking a measurement' involves a scientist using a result in the context of theory to represent a given phenomenon (2017, 9-15). In contrast, 'making a measurement' involves setting up experimental conditions to produce a phenomenon—where that phenomenon can be realized in nature but it can also be a brand new

phenomenon (Keyser 2017, 10). The difference between these two methodologies seems to be a matter of passive representation of a phenomenon vs. active intervention to produce a phenomenon. While the distinction between representation and intervention has been useful in classifying methodology in well-documented contexts like thermometry, microscopy, and cellular measurement, I argue that it falls apart in contexts where taking and making are *entangled*—such as in the context of biomarker measurement in the biomedical sciences.

In this discussion, I aim to show that in *complex methodological contexts*, representational and intervention-based roles require re-conceptualization. I analyze the *relations* between representation and intervention by focusing on the role of intervention in *mediating* representations. In Section 2, I show how applied scientific practice challenges the simple distinction between representational and intervention-based roles of experiment/measurement. In Section 3, I discuss the complex interaction between representation and intervention applied to methodology in biomarker measurement.

2. Methodology at the Intersection between Intervention and Representation

In order to understand why the distinction between representation and intervention needs a multifaceted approach, it is important to be explicit about what it means to represent and intervene in scientific practice. In Section 2.1, I draw on van Fraassen (2008) to discuss representation and both van Fraassen (2008) and Keyser (2017) to discuss intervention. Then in Section 2.2, I show how applied scientific practice challenges the simplistic distinction between representational and intervention-based

roles of experiment/measurement. I argue that the distinction between intervention and representation is less about *specific types of methodologies* in measurement/experiment and more about where one philosophically partitions the measurement *process*.

2.1. Representation and intervention

In experimental and measurement practice, representation has at least three important components: First, instruments or experimental contexts yield measurement values; Second, those values can only be interpreted within the context of a well-developed theory; and third, the relation between the measurement values and the phenomenon is determined by a user (e.g., experimenter). Van Fraassen (2008) provides a rich characterization of representation in measurement and experiment, which requires careful analysis. Worth noting is that van Fraassen takes measurements to be a "special elements of the experimental procedure" (2008, 93-94). For my discussion the embeddedness of measurement in experiment is not important. I will focus on the roles or processes within measurement and experimental practice. But to do this, I will sometimes refer to 'measurement' and other times to 'experiment'. Van Fraassen's characterization focuses on interaction and representation in measurement:

> A measurement is a physical interaction, set up by agents, in a way that allows them to gather information. The outcome of a measurement provides a representation of the entity (object, event, process) measured, selectively, by displaying values of some physical parameters that—according to the theory governing this context—characterize that object. (2008, 179-180)

For van Fraassen, measurement interaction between an object of measurement and apparatus generates a physical outcome—the "measurement outcome" or "physical correlate of the measurement outcome"—, which provides information content about the target of measurement (2008, 143). The contents of measurement outcomes convey information about *what is measured* through the mediation of theory. Van Fraassen posits that theoretical characterization of measurement interaction requires 'coherence':

> The theoretical characterization of the measurement situations is required to be coherent with the claims about the existence of measurement outcomes, their relation to what is measured, and their function as sources of information. (2008, 145)

In short, the theory tells a coherence story about "how its outcomes provide information about what is being measured" (145). Furthermore, the information content is representational. Van Fraassen says, "The outcome provides a representation *of* the measured item, but also represents it *as* thus or so" (2008, 180). To understand how the representational relation works, it is important to refer to van Fraassen's 'representation criterion':

> The criterion for what sorts of interactions can be measurements will be, roughly speaking, that the outcome must represent the target in a certain fashion—, selectively resembling it at a certain level of abstraction, according to the theory—*it is a representation criterion*. (van Fraassen 2008, 141).

Two aspects of the representation criterion require explanation: First, the distinction between "target" and "outcome"; and second, the role of theory in the operation of measurement. I begin with the former. Van Fraassen makes a technical

distinction between the target of measurement ('phenomena') and the outcome of measurement ('appearances'):

> Phenomena are observable, but their appearance, that is to say, *what they look like in given measurement or observation set-ups*, is to be distinguished from them as much as any person's appearance is to be distinguished from that person. (2008, 285)

For van Fraassen, phenomena are observable objects, events, and processes (2008, 283). He emphasizes that phenomena include all observable entities—whether observed or not (2008, 307). A given phenomenon can be measured in many different ways. The outcome of each measurement provides a perspective on a given phenomenon—meaning that the content of measurement tells us what things *look like*, not what they *are like* (2008, 176, 182). The *content* of the measurement outcome is an appearance.

An important qualification is that for van Fraassen, a representation does not represent on its own. The scientist selects the aspects/respects and degrees to which a representation represents a target. This relation can be expressed as: Z uses X to represent Y as F, for purposes P.

Now that the target and outcome of measurement have been characterized, we can specify van Fraassen's role of theory in measurement. According to van Fraassen, "Measurement is an operation that locates an item (already classified as in the domain of a given theory) in a logical space, provided by the theory to represent a range of possible states or characteristics of such items (164). Three things are worth noting about van Fraassen's discussion of logical spaces. First, a logical space provides a multidimensional mathematical space that locates potential objects of measurement (2008, 164). By

measuring we assign the item a location in a logical space. However, according to van Fraassen, it does not have to be on a real number continuum. As van Fraassen points out, items may be classified (by theory) on a range that is "an algebra", "lattice", or a "rudimentary poset" (2008, 172). Second, theoretical location depends on a "family of models" and not just an individual model (2008, 164). Third, an item is located in a "region" of logical space rather than at an exact point (2008, 165). Simply put, theory provides a classificatory system for what is measured. Importantly, theory is *necessary* for this type of classification. Van Fraassen says, "A claim of the form "This is an X-measurement of quantity M pertaining to S" makes sense *only* in a context where the object measured is already classified as a system characterized by quantity M" (2008, 144 my emphasis).

We can summarize the above discussion into four conditions for van Fraassen's account of representation in measurement/experiment practice:

i. Physical Interaction Condition: The interaction between apparatus an object produces a physical correlate of the measurement outcome.

ii. Theoretical Characterization Condition: The content of the measurement outcome is given a location in a logical space, which is governed by a family of theoretical models. An item's location within a logical space can change in content and truth conditions as accepted theories change.

iii. Representational Content Condition: The content of a measurement outcome provides a selective representation of a given target of measurement (phenomenon). Because representations do not represent on their own, users and pragmatic considerations set the representational relation such that: Z uses X to represent Y as F, for purposes P.

iv. Perspectival Information Condition: Measurement generates appearances, which are public, intersubjective, contents of measurement outcomes. Appearances provide selective information about phenomena. Thus information from measurement tells us what something *looks* like and not what something *is* like.

Van Fraassen notes that measurement and experiment are not only limited to a representational role, they can take on at least two productive roles. First, instruments can produce phenomena that "imitate" natural phenomena. That is, carefully controlled conditions give rise to mimetic effects that are used by scientists in the context of theory to resemble natural phenomena (2008, 94-95). It is important to note that van Fraassen emphasizes that natural phenomena are phenomena that exist *independent of human intervention* (2008, 95). The second productive role of instruments is that they are used as "engines of creation" to produce or manufacture new phenomena. Van Fraassen is not explicit about whether or not the representational roles can smear with the productive roles. There is no reason to assume that these roles cannot be combined; but that requires explicit philosophical work to see *how*, which I develop in Section 3.

Keyser (2017) is explicit about the relationship between the representational and intervention-based roles in science. He discusses the *use* of intervention for developing causal representations. Scientists intervene, thereby manipulating causal conditions within a given measurement or experimental system, which he calls 'intervention systems', to produce some sort of "effect" (Keyser 2017, 9-10). According to Keyser, "Intervention systems consist of organized experimental conditions and as such the effects that emerge are often sensitive to changes in conditions*"* (Keyser 2017, 10). Once a given effect is produced it can be used in order to be informative about causal relations for theoretical model building.

Keyser (2017) also differentiates between the methodologies of taking measurements vs. making measurements. I interpret that taking measurements involves three components: First, some instrument or experimental arrangement yields a qualitative or quantitative value; second, a 'theoretical representational framework'—which is just a body of models—is necessary in order to characterize that value according to parameters and relations between parameters; and third, a scientist sets up the resemblance relation between the measurement/experiment value and some aspect(s) of a phenomenon (Keyser 2017, 14-15). In contrast, when scientists make measurements they manipulate causal conditions—such as, preparatory, instrument, and background conditions—within an intervention system. This manipulation gives rise to some effect (Keyser 2017, 3-12).

There is something puzzling about Keyser's distinction between making vs. taking, if we apply the aforementioned conditions (i-iv): i. *Physical Interaction Condition*; ii. *Theoretical Characterization Condition*; iii. *Representational Content*

Condition; and iv. *Perspectival Information Condition*. Namely, it seems that 'making measurements' is compatible with conditions i-iv, so it is not clear why there is a need for a distinction in methodological type, but rather just a difference in details for each condition. For example, when a measurement is made, there is a (i) *physical interaction* that occurs, but it is broader than just the instrument and object. The interaction can include "experimental conditions" (Keyser 2017, 3-5). The product of a made measurement is also amenable to (ii) *theoretical characterization*. Keyser emphasizes that theoretical characterization is necessary for experiment/measurement (Keyser 2017, 14); but he does not make the additional move to say that theoretical characterization is *part of the process* of making a measurement. That is, in order to make a measurement about an effect, one needs to also *characterize* that effect. Without the final characterization, one is only dealing with the material conditions, which is an incomplete part of the measurement process. Keyser can accept that theoretical characterization is a necessary component of making a measurement. Otherwise, he risks offering a limited concept of 'making a measurement' that only applies to arranging the material components of the measurement process and nothing further.

The same challenge goes for (iii) *representational content* and (iv) *perspectival information*. An important component of the measurement process is to represent the relation between the produced effect and some aspect(s) of a phenomenon. For example, is this given effect a limited mimetic representation of a natural phenomenon or is it a brand new phenomenon? Without claims about what the effect is and its relation to objects, events, and processes in the world, 'making a measurement' is uninformative about part of the measurement process: the final value of the measurement outcome.

The aforementioned considerations question the need for a distinction between 'making' vs. 'taking'. One conclusion is that making uses the same components (i-iv), just with slightly different detail. But the other conclusion is a bit unsatisfying: making is really only about organizing the material components, which is an *initial* step in the measurement process, and it does not apply to later steps in measurement.

2.2. Dynamic relations between intervention and representation

I argue that the distinction between intervention vs. representation is less about *specific types of methodologies* in measurement/experiment and more about where to philosophically partition the *measurement process*. To make this point clear, I make two sub-points: 1) Measurement in the biological sciences offers complex and sometimes blurred relations between instrument and object of measurement such that representation and production take on dynamic roles; 2) There is a difference between the act of measurement and the total process of measurement. I briefly describe (1) and (2).

On van Fraassen's (2008) and Keyser's (2017) characterizations of *representation* in measurement, the role of the instrument/apparatus seems to have an important mediating function. It may be the case that philosophical focus on case studies (e.g., thermometry, microscopy, cellular bio, and bacteria) that are instrument-intensive provide a certain support for an instrument-centric account of representation in measurement. Whether or not the necessary mediating role of instruments is an explicit part of both accounts, there is room to develop a richer philosophical view of the role of representation in the total measurement *process*. Without such philosophical development, we risk missing complex cases of measurement where intervention occurs

side-by-side with representation. For example, in some cases of biological measurement, scientists use the organism to measure processes in that same organism but also to represent larger phenomena (Prasolova et al. 2006). For example, mouse diets are manipulated in order to measure chromatin pattern changes. I characterize this as the mouse *constituting experimental conditions* that are being manipulated in order to measure some sort of process. The manipulation of conditions indicates an interventionist approach (or 'making' a measurement). Moreover, without manipulating the mouse's diet scientists would not be able to make a reliable measurement on chromatin structure at all. So the organism is not only being manipulated as part of the experimental/measurement set-up, it is a crucial part of that set-up. That is, without intervention, there is no reliable result. In addition to the organism being used as part of the measurement set-up, it also serves as a physical *representation* of the dynamics of chromatin pattern change. That is, a given model organism can serve as a data model for a specific phenomenon of study—e.g., chromatin pattern in organism X. So, in this case the organism serves a dual function: it constitutes a set of experimental conditions to be manipulated and it serves as a physical representation of a phenomenon. Because of the dual function, this seems to be a case of both 'making' and 'taking' a measurement.

This brings me to sub-point (2). The total process of measurement is often complex in the biological sciences and requires multiple stages of intervening and representing. As mentioned in the model organism example representation and intervention are often *entangled*. Measurement is not merely putting an instrument up to something and waiting for a reading, which can be classified as an *act* of measurement. Measurement is also not merely creating effects out of material conditions. Measurement

requires manipulation of conditions that is *used* in order to generate a representation. For example, identifying a mysterious fungus that is entangled with other fungus in a sample is an active process that requires both intervention and representation. One method is to take a sample and scrape it over a petri dish. What grows are spores that are passively deposited. But if common fungi were commingled with the mysterious fungi in the sample, and the common fungi grew faster, it would be impossible to identify the mysterious fungus. That is, coming back in a couple of weeks and seeing the petri dish covered with familiar species would lead to a false conclusion. Another way to perform the measurement (i.e. culture samples) is as follows. Take the samples and grind them up. Then sprinkle them into a petri dish. Put the dish under the microscope and, using a fine needle, pick out fragments of the mysterious fungus and transplant them to their own dishes (Scott 2010). Once the fragments have been transplanted through this fine-grained intervention, each dish can be left to grow the colonies. The final dishes will offer visual representations that serve as data on the nature of the mysterious fungus. Notice here that intervention is a precursor to reliable representation.

Representation is not only reserved for the final instrument reading. It can also occur at other stages in the measurement process. Likewise, manipulation does not have to occur only at the earlier stages. For instance, organic matter can function as an instrument, like in the case of FourU thermometers, which are RNA molecules that act as thermometers in Salmonella (see Waldminghaus et al. 2007). Suppose that a scientist sets up an experiment to iteratively measure to what extent modifying RNA factors in FourU thermometers changes thermometer readings in Salmonella. In such a case the scientist could modify molecular factors and use the organic thermometers as temperature

measures over many iterations, which would culminate in some sort of data model that organizes the relationship between molecular factors and FourU function. In such a case, there are multiple layers of intervention and representation.

The complex layering of intervention and representation is apparent in biomarker measurement in the biomedical sciences, where biological components serve as representations of disease conditions, but are also intervened on in order to make more reliable representations. I turn to this case study in the subsequent section.

3. Intervening in Representations and Representing Interventions

Biomarkers are used in biomedical measurement to reliably predict causal information about patient outcomes while minimizing the complexity of measurement, resources, and invasiveness. A biomarker is an assayable metric—or simply, an indicator—that is used by scientists to draw conclusions about a biological process (De Gruttola et al. 2001). The greatest utility from biomarker measurement comes from their ability to help clinicians and researchers make conclusions with limited invasiveness. The reliance on biomarkers to make causal conclusions has prompted the use of 'surrogate markers'. These biomarkers are used to substitute for a clinically meaningful endpoint such as a disease condition. A major scientific methodological issue is that the use of multiple biomarkers will produce disagreeing results—and this is true even in the context of biomarkers that use similar biological pathways. To make methodological matters worse, theoretical representation is often not equipped to fill in the causal detail for each biomarker measurement. This amounts to an unfolding methodological puzzle about how to use intervention and representation in biomarkers to produce reliable measurements.

My interest in this case study is not in solving the methodological puzzle, but rather in showing the *relations between intervention and representation* in such a complex case study. In this section, I discuss the complexity of intervention and representation in biomarker measurement to illustrate how intervention mediates the measurement process.

To understand the complex methodology in biomarker measurement it is important to detail the use and limitations of biomarkers. Some biomarkers are used as a substitute for some clinical endpoint. For instance, LDL cholesterol (LDL-C) is a biomarker that clinicians and physicians use to correspond to a clinical endpoint—e.g., heart attack. Moreover, the biomarker is associated with risk factors such as coronary artery stenosis, atherosclerosis, and angina pectoris. Katz (2004) argues that all biomarkers are candidates for 'surrogate markers', which can serve as substitutes for clinical endpoints. That is, surrogate markers are reliable biomarkers that have a one-to-one correspondence with the disease condition such that they can be used to provide reliable predictive and causal information about a given clinical endpoint. There are a couple of points worth noting. First, notice that biomarkers and surrogate markers are being used as representations of a clinical endpoint. That is, to figure out the likelihood of developing a disease condition and to understand the risk factors associated with that disease condition, scientists use biomarkers that indicate information about the endpoint. This means that these physiological components can be used by clinicians and physicians to *represent disease conditions to respects and degrees*. The second point worth noting is that there are many biomarkers but limited surrogate markers and even more limited validated surrogate markers ('surrogate endpoints')—which are surrogate markers that are reliable in multiple contexts of interventions. The importance of this will be relevant

shortly when I discuss the complexity of biomarker measurement. For our purposes, this means that most biomarkers in biomedical practice provide very limited representational information.

Surrogate markers are not passively used as physical representations of disease conditions. Their use is often more effective for representational purposes if there is a *mediating intervention*. For instance, surrogate markers can constitute "response variables". This is where a surrogate marker is manipulated in order to produce an effect that is relevantly similar to the effect with the same manipulation on the clinical endpoint. This means that an adequate surrogate must be "tightly correlated" with the true clinical endpoint; but it also means that any intervention on a surrogate marker must be tightly correlated with the intervention on the true clinical endpoint (Buyse et al. 2000). I interpret this as a dual role for a reliable surrogate marker. It is to act as an epidemiological marker that *represents* some clinical endpoint but also to act as a responding variable that can be used in an *intervention* to causally influence the clinical endpoint. An example of the dual role of the surrogate marker is that high concentrations of LDL cholesterol (LDL-C) correspond to cardiovascular risk (Gofman and Lindgren 1950). But if a therapeutic interventions is used—such as, 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase inhibitors (statins)—that intervention can lower LDL levels, which in turn reduces cardiovascular disease (LaRosa et al. 2005).

So far I have presented the representational and intervention-based role of biomarkers. It is not straightforward to say that surrogate markers are *'made'* like an effect. But it is also not straightforward to say that surrogate markers constitute a *measurement outcome that is the final reading on an instrument*. These markers provide

useful representational information *in the context* of an intervention. To add to the complexity of the relation between representation and intervention, biomarkers in the context of Alzheimer's measurement have added methodological steps. In Alzheimer's measurement there are different biomarkers, which are not correlated with each other and change with independent dynamics in the progression of Alzheimer's disease. So *each* of these biomarkers do not provide the same type of representation about the progression of Alzheimer's disease. Furthermore, scientists *only* understand the disagreement between each of these biomarkers in the presence of different interventions.[1] The different interventions are in the form of drugs (e.g., bapineuzumab and solanezumab) and these interventions produce disagreeing representational results for the biomarkers. That is, the biomarkers respond differently to different interventions, which is methodologically problematic because it indicates that all of these biomarkers cannot be reliably tracking Alzheimer's progression in the same way. Interestingly, scientists systematically compare these disagreeing results to make reliable claims about Alzheimer's progression and treatment (Toyn 2015).[2] To simplify the method used, scientists track how interventions

[1] There has been much work recently on clinical biomarkers like: cerebrospinal fluid (CSF) tau, which is the primary component of neurofibrillary tangles; CSF 42-amino acid amyloid-β (CSF Aβ), which is the protein cleavage product believed to precipitate disease by forming neuron-damaging plaques; and amyloid plaques from PET scans. While the methodological story is beyond the scope of this discussion, there is a complex methodological point that is noteworthy for this discussion (Toyn 2015).

[2] To give a brief picture: The intervention of Bapineuzumab reduces levels of plaque assayed by Aβ PET and CSF tau, but not CSF Aβ; but Solanezumab *does not alter* levels

change properties of biomarkers and then they compare these amalgamated results with how interventions change behavioral/cognitive properties. This type of cross comparison allows scientists to eliminate biomarkers that do not track behavioral/cognitive improvement.

The structure of the methodological complexity in biomarker measurement can be partitioned as follows: 1) For a particular clinical endpoint, there are *limited physical representations* in the form biomarkers (or surrogate markers) which can be *used* to make representational and perspectival conclusions about the endpoint or risk factors associated with it; 2) *Scientists intervene in a process* from each of the biomarkers in order to track the relations between biomarkers and clinical endpoints; and 3) Such interventions prompt *disagreeing results between the biomarkers*, which can 4) be amalgamated by researchers into further representations of the *relations between biomarkers and their clinical endpoints*. The above structural breakdown is merely *a* type of complex methodological process that can occur in biomedical measurement. It shows how interventions on physical representations (biomarkers) can produce other reliable representations. What is important to note about this analysis is the role of intervention in *mediating* further representations. In the case of biomarkers, intervention is necessary to test how close biomarkers are in their representations of clinical endpoints and also to other biomarkers. These representations not only represent the relation between the original biomarker and the clinical endpoint, but they also represent how a given

---

of plaque assayed by Aβ PET and CSF tau but leads to a *reduction in* CSF Aβ. Cross comparison of the *intervention* mechanisms allows scientists to begin to make causal claims about which biomarkers are more reliable than others (Toyn 2015).

intervention affects a given biomarker. As such, intervention paves the way for iterations of representations.

4. Concluding Remarks

In this discussion, I have analyzed the role of intervention in mediating representations by using examples from the biological and biomedical sciences. Characterizing intervention as a mediating factor in a larger methodological operation provides an important point about scientific practice. Representation and intervention are not neatly partitioned into contrasting methodologies. In fact, applied science often dictates the complex, and often smeared, philosophical concepts and methodologies. For this reason, I am proposing a *process* view of intervention and representation. This view opens up the diversity of relations between representation and intervention in a given experimental/measurement practice. While I have emphasized how intervention mediates representation, there is more territory to explore about the mediating role of representation for intervention.

Work Cited

De Gruttola, V.G, Clax P, DeMets DL, et al. (2001). Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. Control Clin Trials 22:485–502.

Gofman, J.W., Jones, H.B., Lindgren, F.T., et al (1950). Blood lipids and human atherosclerosis. *Circulation* 2:161–178.

Hacking, I., (1983). *Representing and Intervening*, Cambridge: Cambridge University

Press.

Heidelberger, M. (2003). Theory-ladenness and scientific instruments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 138–151). Pittsburgh, PA: University of Pittsburgh Press.

Katz, R. (2004). Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* 1:189–195. doi: 10.1602/neurorx.1.2.189

Keyser, V. (2017). Experimental Effects and Causal Representations. *Synthese*, SI: Modeling and Representation, pp. 1-32.

LaRosa, J.C., Grundy, S.M., Waters, D.D., et al. (2005). Intensive Lipid Lowering with Atorvastatin in Patients with Stable Coronary Disease. *New England Journal of Medicine* 352:1425–1435. doi: 10.1056/NEJMoa050461

Prasolova L.A., L.N. Trut, I.N. Os'kina, R.G. Gulevich, I.Z. Pliusnina, E.B. Vsevolodov, I.F. Latypov. (2006). The effect of methyl-containing supplements during pregnancy on the phenotypic modification of offspring hair color in rats. *Genetika*, 42*(1)*, 78-83.

Radder, H. (2003). Technology and theory in experimental science. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 174–197). Pittsburgh, PA: University of Pittsburgh Press.

Toyn, J. (2015). What lessons can be learned from failed Alzheimer's disease trials? *Expert Rev Clin Pharmacol* 8:267–269. doi: 10.1586/17512433.2015.1034690

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.

Waldminghaus, T., Nadja H., Sabine B., and Franz N. (2007). FourU: A Novel Type of

RNA Thermometer in Salmonella. *Molecular Microbiology* 65 (2): 413–24. https://doi.org/10.1111/j.1365-2958.2007.05794.x.

Philosophy of Science (forthcoming)
v1.2 (as of 9/15/18)
Please cite published version

Are Emotions Psychological Constructions?

Charlie Kurth
Department of Philosophy
Western Michigan University

Abstract: According to psychological constructivism, emotions result from projecting folk emotion concepts onto felt affective episodes (e.g., Barrett 2017, LeDoux 2015, Russell 2004). Moreover, while constructivists acknowledge there's a biological dimension to emotion, they deny that emotions are (or involve) affect programs. So they also deny that emotions are natural kinds. However, the essential role constructivism gives to felt experience and folk concepts leads to an account that's extensionally inadequate and functionally inaccurate. Moreover, biologically-oriented proposals that reject these commitments are not similarly encumbered. Recognizing this has two implications: biological mechanisms are more central to emotion than constructivism allows, and the conclusion that emotions aren't natural kinds is premature.

This paper challenges the psychological constructivist account of emotions that is gaining prominence among neuroscientists and psychologists (e.g., Barrett 2017, 2012, 2009; LeDoux 2015; Russell 2004). According to constructivism, emotions result from projecting culturally-fashioned concepts onto felt affective episodes. Fear, for instance, just is a feeling of negative arousal as viewed through the lens of one's folk concept FEAR. This proposal is novel in taking felt experience and cognitive projection to be essential elements of what emotions are. Moreover, while constructivists acknowledge that there's a biological dimension to emotions (e.g., neural mechanisms are responsible for generating the conscious feelings that we project our emotion concepts on to), they deny that emotions are, or necessarily involve, anything like an affect program. Thus, constructivism is philosophically significant in two ways. First, in denying an essential role for biological mechanisms, it challenges influential, affect-program-oriented accounts of emotion (e.g., Scarantino & Griffiths 2011; Ekman & Cordaro 2011). Second, in understanding emotions as projections of folk emotion concepts, it takes emotions to be social-psychological constructions, not natural kinds.

But despite constructivism's appeal among cognitive scientists, the role that it gives to felt experience and folk concepts leads to an account of emotion that's both extensionally inadequate and functionally inaccurate. Moreover, biologically-oriented proposals that reject constructivism's problematic commitments are not similarly encumbered. Recognizing all this reveals that an adequate account needs to give greater place to the biological mechanisms that underlie emotions than constructivism allows. This, in turn, suggests that the constructivists' conclusion that emotions are not natural kinds is premature.

1. Psychological Constructivism and Its Appeal

Constructivism sees emotions as having two elements: a felt affective experience and a cognitive projection or labeling. Taking these in turn, the felt experience component—or "core affect" as it's often called—is a neurophysiological state that manifests as a consciously experienced combination of valence (i.e., feeling good or bad) and arousal (i.e., feeling activated or deactivated) (Barret 2006: 48; Russell 2004; LeDoux 2015: 226-232). Importantly, constructivism's focus on core affect looks just to the amalgamated *experience* of these two components—valence and arousal. What *causes* this felt experience is irrelevant to the nature and individuation of emotions. In fact, and as we will see, allowing that particular sensations (instances of core affect) can be produced by a range of distinct neural circuits or somatic events is taken to be a point in favor of the constructivist proposal.

Given this account of the felt dimension, constructivism maintains that "discrete emotions emerge from a conceptual analysis of core affect. Specifically, the experience of feeling an emotion…occurs when conceptual knowledge about emotion is brought to bear to categorize a momentary state of core affect. … [These] [c]ategorization processes enact the rules, [that guide] the emergence of an emotional episode" (Barrett 2006: 49; also LeDoux 2015: 225-232). This talk of "conceptual analysis," "conceptual knowledge," and "categorization" should be understood thinly.

The underlying process needn't involve some full-fledged, conscious judgment. Rather, all that's necessary is an unconscious or implicit recognition that one's sense of one's situation, and one's felt physiological state, fall under a particular folk emotion concept.

These emotions concepts, in turn, should be understood as folk theories or culturally-shaped behavioral scripts that detail the nature and function of the particular mental states picked out by specific emotion labels ('fear,' 'joy,' 'anger,' etc.). Moreover, the fact that folk emotion concepts engage these folk theories and behavioral scripts entails that the projecting of a particular label onto an instance of core affect not only imbues one's situation with the associated, emotionally-colored meaning, but also shapes one's subsequent thoughts, physiological responses, and behaviors (Barrett 2012; LeDoux 2015).

Formalizing this a bit, we can see psychological constructivism as committed to four theses:

> (PC1) Each emotion type/category is constituted by the projecting of a specific folk emotion concept (e.g., FEAR, JOY) onto a felt affective experience.
>
> (PC2) Token emotion episodes (e.g., a given instance of fear) are cognitive acts where one (implicitly) labels an occurrent conscious feeling with a particular folk emotion concept and so comes to see the feeling through the lens of that concept.
>
> (PC3) There is no unique (set of) neural circuit(s) or psychological mechanism(s) responsible for the conscious feelings that get categorized with particular folk emotion concepts.
>
> (PC4) The act of labeling a feeling with a particular folk emotion concept affects one's subsequent thoughts, physiological responses, and behaviors.

According to its advocates, much of constructivism's appeal lies in its explanatory power. In comparison to more biologically-oriented theories, it provides a better explanation of empirical research on the biological mechanisms and correlates associated with emotions (e.g., neural circuits, patterns of physiological change, and expressive behavior). Since the discussion that follows will build

from the contrast between constructivism and competing biologically-oriented theories (BTs), it will be useful to briefly sketch the BT approach and the constructivists' case against it.

As a generalization, BTs maintain that emotions are, or necessarily engage, affect programs—that is, largely encapsulated systems that automatically prompt stereotyped patterns of physiological changes, expressive behavior, motor routines, attentional shifts, and forms of higher-cognitive processing in response to (evolutionarily-relevant) threats and opportunities. So, for example, fear is (or essentially involves) an affect state that consists of automatically engaged tendencies for inter alia increases in arousal, narrowing of attention, and the cueing of fight/flight/freeze behavior in response to the perception of some danger.

But since BTs take affect programs to be essential (even identical) to emotions, constructivists argue they cannot explain two well-documented sets of findings.[1]

> (F1) One can feel a given emotion without engaging what science suggests is the best candidate for its underlying biological drivers (or their correlates)—e.g., activation of particular neural circuits, a distinctive physiological response, characteristic expressive behavior.
>
> (F2) The relevant biological drivers/correlates can be engaged though one does not report feeling the associated emotion.

So, for instance, though the central nucleus of the amygdala (CeA) is thought to be central to fear, research shows both that individuals will report being afraid when the CeA is not engaged (F1), and that the CeA can be active though individuals report not feeling fear (F2).

BT proponents have sought to address these explanatory limitations by insisting that we must narrow our understanding of what, say, FEAR is. More specifically, they maintain that the folk emotion concepts that the above research relies on (in, e.g., the self-reports of emotions (not) felt) are too

[1] See, e.g., Barrett 2012 for a review of the relevant empirical work.

coarsely grained for scientific investigations like these. The BT advocates' expectation is that a more refined account of what 'fear' refers to will reduce, even eliminate, dissociations of the sort noted above (e.g., Scarantino & Griffiths 2011; Kurth 2018). But constructivists respond that any effort to narrow or otherwise refine our emotion concepts along these lines will result in an account of (e.g.) fear that is troublingly stipulative or excessively revisionary with regard to our ordinary understanding of these emotions (Barrett 2012: 415-6; LeDoux 2015: 234).

Two aspects of this debates are particularly important for our purposes. First, central to the constructivist complaint is the move to take a failure to accommodate our *ordinary emotion talk* as the standard for what counts as stipulative or excessively revisionary account. Second, given our ordinary emotion talk as the standard, the above four theses appear to give constructivism the resources and flexibility it needs to explain not just (F1)-(F2), but also the richness and cultural variation of emotional life more generally (e.g., Barrett 2012, 2009). However, I will argue that investigating the extensional adequacy and functional accuracy of constructivism's core theses provides us with reason to doubt each of (PC1)-(PC4).

2. Is Constructivism Extensionally Adequate?

As we've seen, a central feature of the debate between constructivism and BTs is the charge that BTs cannot accommodate dissociation data without committing to a stipulative or excessively revisionary account of what emotions are. In what follows, I give three examples that suggest constructivism faces a similar problem. More specifically, a closer look at the constructivists' dual claim that emotions are *cognitive labelings* of *felt experiences* reveals that the account is both under- and over-inclusive with regard

to our ordinary understanding of things like: what emotions are, when we experience them, and how they differ from moods, feelings, and other categories of affect.[2]

First consider the constructivist's commitment to understanding emotions as felt experiences—that is, changes in core affect that we're consciously aware of. An implication of taking felt affective experience as essential to being an emotion is that it rules out the possibility of unconscious emotions. Some constructivists appear to embrace this result. For instance, Joseph LeDoux maintains that claims about unconscious emotions are "oxymoronic" (2015: 234; also, 19). But LeDoux's acceptance of this implication aside, the thought that there cannot be unconscious emotions fits poorly with our everyday experiences and our ordinary emotion talk.

For instance, if there aren't unconscious emotions, then how do we explain situations where we don't realize that we were (say) afraid until *after* the danger has passed? Pressing further, notice that we not only regularly speak of unconscious emotions, but also appeal to them in order to explain our behavior. For example, we say things like, "Bill won't discuss the book he is working on. He says it's not ready yet—but he doesn't realize that he's really just afraid about getting negative feedback." While ordinary talk like this is easy to make sense of on the assumption that Bill is unconsciously fearful, such an explanation isn't available to a constructivist like LeDoux—our ordinary talk to the contrary, Bill isn't unconsciously afraid, but rather experiencing some other psychological blockage.

But the constructivists' trouble with unconscious emotions runs deeper—the case for their existence also has empirical support. For instance, recent experimental work has shown that subliminally presented emotion faces can produce affective responses that bring emotion-specific behaviors *even though* the subject denies feeling an emotion. In particular, subliminally presented happy

[2] Thus the strategy I employ here—one that *grants* constructivists' their criterion for assessing when an account is excessively revisionary—is distinct from standard defenses of BTs noted in §1.

faces bring increased "liking" behavior (e.g., greater consumption of a novel beverage), while subliminally presented angry faces have the opposite result (Winkielman et al. 2003; also, Kihlstrom 1999). Since these patterns of behavior mesh with our understanding of both joy as an emotion that tends to increase interest/engagement, and anger as an emotion that brings avoidance/rejection tendencies, these results are taken as evidence of unconscious emotions.

While the constructivist might try to pass these findings off as cases where unconscious changes in core affect (not emotion) produce the behaviors, the plausibility of the proposal is undercut by the fit we find between the subliminally presented happy (angry) face, the resulting liking (avoidance) behavior, and *our ordinary understanding* what happiness (anger) involves (Winkielman et al. 2005). The upshot, then, is that constructivism's insistence that felt changes in core affect are *essential* to what emotions are has revisionary implications with regard to our ordinary (and scientific) understanding of emotional life.

But even if we're willing to grant that our talk of unconscious emotions is merely metaphorical—an elliptical way of talking about some non-emotion form of (unconscious) affect—the constructivist's second core commitment brings additional problems. In particular, the claim that emotions are the product of our cognitive labelings/projections makes facts about when we are experiencing an emotion—and what emotion it is—too sensitive to random situational features and framing effects. To draw this out, consider the following case.

> Coffee. I order a cup of decaf coffee and sit down to read a magazine cover story about Trump's latest foreign policy provocations. But unbeknownst to me, the barista confuses my order and I get a cup of regular coffee. As the caffeine works its way into my system, it brings a (consciously experienced) change in my arousal. As a result, I start reading the article with jittery attentiveness.

Given the scenario, it seems my jittery, attentive reading is best understood as a bout of caffeine-induced hyperactivity. But notice: there's nothing in the constructivist account to rule out the

possibility that I'm actually having an emotional experience—I'm afraid. After all, on the constructivist account, this experience could be a change in core affect that I've (implicitly) labeled 'fear.' While that possibility alone seems odd (to my ear, at least, the case is best understood as emotionless hyperactivity, not fear), there's more trouble.

To draw this out, consider the constructivist's likely response to the case. Given the setup, she would likely maintain that whether this is an instance of fear depends on whether I see it that way—what sort of meaning do I attribute to my situation (e.g., Barrett 2017: 126; 2012: 419-420; 2009: 1293)? For instance, if I assent to the barista's remark that I seem really uneasy about the article that I'm reading, then—by (implicitly) labeling my behavior through my assent—I imbue my situation with the meaning carried by my FEAR concept. I am, therefore, feeling fear. While this move might seem to allow the constructivist a way to account for the case, it comes at a high cost. For notice, had the barista instead said something like, "Whoops, I messed up and gave you regular, not decaf—no wonder you're so hyper," I'd likely assent to that too. And so I wouldn't be afraid—just hyperactively aroused.

But that's odd. Our ordinary thinking about emotions suggests that whether I'm experiencing a particular emotion, and what emotion I'm experiencing, should *not* be so sensitive to random situational features like what questions the barista—or anyone for that matter—just happen to ask me. To be clear, the claim here is not that emotions are immune to situational and contextual factors. Rather, the point is that on the constructivists' account emotions turn out to be *too* sensitive to them. The radical situational sensitivity entailed by constructivism makes it not only too easy to experience an emotion, but also ties facts about what emotion we're experiencing to irrelevant situational factors.

Together, the difficulties raised by unconscious emotions and incidental situational features call the extensional adequacy of the constructivist account into question and do so in a way that

pinpoints the commitments of (PC1) and (PC2) as the source of the trouble—after all, these claims posit feelings of core affect and projections of folk concepts as essential to what emotions are. Of equal note is the fact that biological theories are less vulnerable to these difficulties. For one, irrelevant situational features should have less influence on what emotion one happens to experience since, according to BTs, emotions are (or are principally driven by) affect programs, not contextualized cognitive labelings. Moreover, since affect programs are things that can operate below that level of conscious awareness (Kurth 2018), taking emotions to be driven by affect programs provides BTs with the resources needed to explain unconscious emotions.

While the above discussion raises worries about the first two constructivist theses (PC1-PC2), it also provides the makings for worries about the third. In particular, because constructivism denies (via PC3) that emotions are underwritten by affect programs, it has trouble making plausible distinctions between emotions and similar states like moods. To draw this out, notice that the coffee case from above can be easily extended to show that constructivism makes it too easy to flip between moods and emotions. All we need to do is substitute "being in a worried mood" for "hyperactive" in the presentation of the case. Once we do this, we see that mere changes in the question the barista asks me can change whether I'm worried (a mood) or afraid (an emotion).

So we again see that constructivism has problematic explanatory limitations—this time with regard to preserving the thought that there's a substantive difference between moods and emotions. On the constructivist account, this distinction is just a matter of how we happen to label our felt experiences. While some constructivists appear willing to accept this conclusion (e.g., Barrett 2017, 2009), it highlights another place where the constructivist proposal has revisionary implications—after all, moods and emotions are generally thought to be *distinct* forms of affect (e.g., Ben-Ze'ev 2000: Chap. 4). Moreover, here too we have a difficulty that's easily avoided by biological accounts. Since

BTs take emotions to be (driven by) affect programs, they can appeal to the engagement of these mechanisms as the basis for the emotion/mood distinction (e.g., Kurth 2018; Wong 2017).

Stepping back, then, although constructivism purports to be less stipulative with regard to capturing our ordinary understanding of emotions, the above examples call this into question. For starters, the constructivists' commitment to (PC1)-(PC3) has revisionary implications for our ordinary understanding of what emotions are, when we experience them, and how they differ from moods. Moreover, we have also seen that biologically-oriented accounts—in eschewing this trio of problematic theses—are better equipped to provide a plausible account of these features of our everyday emotion talk.

3. Is Constructivism Functionally Accurate?

The challenges to the constructivist picture extend beyond concerns about its extensional adequacy. The account also makes predictions about how projecting emotion concepts onto felt experience should shape subsequent behavior that are poorly supported by the empirical record. Two examples will draw this out.

First consider emotion misattribution research. In this work, a feeling that is typically associated with a particular emotion (e.g., feelings of unease and anxiety) is subtly induced, but the individual is lead to believe they are not, in fact, experiencing that emotion but rather something else (e.g., the effects of caffeine). Constructivism predicts (via PC4) that individuals in these experiments should display different behaviors depending on whether they are in the control or misattribution conditions. For instance, individuals led to believe that the unease they're feeling is not anxiety, but something else (caffeine) should display diminished anxiety-related behaviors in comparison to controls who were not misled about their unease. But on this score, the experimental findings are decidedly mixed.

First, while there is a sizable body of findings showing misattribution manipulations attenuate subsequent emotion-related behavior, there is also a sufficiently large set of non-confirmations to raise concerns. For instance, while some research on public speaking anxiety suggests that attributing unease to a pill you just took rather than anxiety about a public talk you must give leads to a reduction in anxiety-related behaviors—stuttering, apprehension, and the like (Olson 1988), other studies have failed to find any differences in these behaviors (Slivkin & Buss 1984; Singerman, Borkovec & Baron 1976).

Moreover, even in cases where emotion-related behavior is reduced in the manipulation condition, it's not clear how much support this brings to the constructivist. This is because it's often unclear whether the reductions in emotion-specific behavior are (i) the result of the misattribution or (ii) a consequence of directing subjects' attention away from the emotion eliciting stimuli (for a review, see, e.g., Reisenzein 1983). This potential confound is problematic for constructivists since only possibility (i) provides direct support for the claim of (PC4)—namely, that the act of labeling *itself* affects subsequent behavior.

The second problematic set of results comes from work in political science. This research investigates how negative emotions shape public policy decision making among voters (e.g., MacKuen et al. 2010; Brader et al. 2008; Valentino et al. 2008). The core hypothesis of this research is that negative emotions (especially, anger and anxiety) affect subsequent behavior in different ways. In particular, anger—as a response to challenges to what one values—should tend to bring behavior geared toward defending the threatened values. By contrast, since anxiety is a response to uncertainty, it should tend to bring caution and information gathering aimed helping one work through the uncertainty one faces.

To test these predictions, the experimental set up works as follows. First, individuals are asked to read a (fake) news story designed to provoke anger or anxiety by challenging the individuals' pre-existing views about contentious policy issues like immigration, affirmative action, and economic policy. After reading the story, the participants are given the opportunity to use a website containing links to additional information, both for and against, the policy issue at hand. They are also asked how the original news story they read made them feel (e.g., angry, anxious). So by tracking what kinds of information the participants looked at through the website, experimenters can identify differences in how the anger and anxiety provoked by the story shaped subsequent behavior.

In the present context, these experiments allow us to test a pair of predictions that follow from the constructivist theses (PC1) and (PC4):

> (P1) Labeling felt experiences with distinct folk emotion concepts should bring different patterns of behavior.
>
> (P2) The behaviors that result from labeling a felt experience with a particular concept should map to our folk understanding of the emotion in question.[3]

More specifically, given (P1) and (P2), we should see different behaviors based on whether the participants in the experiment label their emotion 'anger' or 'anxiety' (P1). Moreover, the different behaviors should map to the above, ordinary understanding of these emotions—e.g., angry individuals should look for information that helps them defend their preferred policy position, while anxious individuals should engage less in motivated inquiry and more in open-minded forms of investigation (P2).

[3] As evidence of constructivism's commitment to these predictions, consider Lisa Feldman Barrett's comment that "when a person is feeling angry…she has categorized sensations from the body and the world using conceptual knowledge of the category 'anger'. As a result, that person will experience an unpleasant, high arousal state as evidence that someone is offensive. In fear…she will experience the same state as evidence that the world is threatening. And, *either way, the person will behave accordingly*" (2009: 1293, emphasis added).

However, whether we find support for these predictions turns—surprisingly—on what the policy issue used in the experiment was. More specifically, in experiments where the policy question that was challenged by the fake news story concerned immigration, the results fit poorly with constructivism's predictions. That is, participants behaved in the same angry way regardless of whether they reported feeling anger or anxiety (Brader et al. 2008). By contrast, if the policy issue at hand concerned affirmative action or economic policy, the results are more in line with (P1)-(P2): anger and anxiety provoked by the news stories not only brought different patterns of behavior, but the resulting behaviors mesh with our ordinary conception of how these emotions function (MacKuen et al. 2010; Valentino 2008).

While this second set of results might appear to be good news for constructivists, the trouble lies in explaining why we get the different results between the immigration and affirmative action/economic policy experiments. After all, other than the content of the issue at hand, the experimental designs were *identical*. In response, the constructivist might argue that content and context matter (e.g., Barrett 2012, 2009): the similar behaviors that subjects display in the immigration version of the study suggest that the cultural scripts associated with 'anger' and 'anxiety' are highly sensitive to negative stereotypes about minorities. More specifically, the thought would be that there's something about the combination of immigration debates and racial stereotypes that changes the standard behavioral scripts associated with 'anger' and 'anxiety' so that, while they *typically* generate different behaviors, they *now* bring the same ones.

But setting aside concerns about the ad hoc nature of this proposal, without more of a backstory, it's unconvincing. After all, affirmative action debates are *also* framed in racial stereotype provoking ways. So here too we should see anger and anxiety generating similar patterns of behavior. But we don't.

Moreover, notice that, on this front, biological accounts have an easier time explaining the experimental findings. For instance, as one possibility, the BT advocate could argue that only participants in the immigration study are likely to be experiencing *both* anger and anxiety: anger about the harms immigrants will bring and anxiety given their uncertainty about the likelihood of these harms. Given this, the BT advocate could then add two claims about what happens when both these emotions are engaged. First, since anger is a more powerful emotion than anxiety, it tends to win out with regard to shaping individuals' subsequent behavior. Second, given the high degree of overlap in the felt experiences produced by the anger and anxiety affect programs (e.g., both bring increased, negatively valenced arousal), when prompted to state what emotion they are feeling, some subjects happen to interpret their feelings as anger, while others see it as anxiety. Thus, the BT advocate can explain both why we get mixed results when subjects are prompted to state what emotion they are feeling and why, despite these differences in self-reports, the individuals nonetheless respond with behavior characteristic of anger, not anxiety. Moreover, because this proposal allows anger to drive behavior *regardless* of how subjects happen to label it, the explanation is unavailable to constructivists.

All told, we have two independent sets of experimental findings showing (at best) equivocal support for constructivism's predictions about how projecting emotion concepts onto felt experience should shape subsequent behavior. Moreover, we've also learned that more biologically-oriented accounts are better able to handle the experimental findings we've reviewed.

4. Conclusion: Emotions, Biology, and Natural Kinds

As we've seen, constructivism's purported advantage over more biologically-oriented theories lies is its ability to better explain the richness and diversity of emotional life (§1). But we have also seen that a crucial premise in this argument is the move to take accommodating our ordinary emotion talk as the standard for assessing a theory's explanatory power. Not only are there familiar problems for

adopting such a standard (e.g., Scarantino & Griffiths 2011, Kurth 2018), but—even if we accept it—we've learned that there's trouble for constructivism. In particular, the explanatory "success" constructivism secures come by way of a highly revisionary account of what emotions are, when we experience them, how they differ from moods, and the way that they shape behavior (§§2-3). Moreover, our critical observations also implicate the four constructivist theses (PC1-PC4) as the source of these difficulties. Thus it's not surprising that more biologically oriented proposals—accounts that reject these commitments—do not face similar explanatory limitations.

Taken together, then, the arguments of this paper suggest a pair of larger lessons. First, even if we agree that constructivists are correct about what the relevant standard for assessing a theory of emotion is, we've learned that an adequate account must give greater place to the biological mechanisms that underlie emotions than constructivism allows. This, in turn, indicates that the constructivists' conclusion that emotions are not natural kinds is premature. After all, if we must posit something like an affect program in order to (i) explain everyday talk and empirical findings about unconscious emotions, (ii) capture the thought that emotional experience is not radically sensitive to random situational features, and (iii) accommodate research regarding how emotions shape behavior, then we have evidence that (at least some) emotions are underwritten by mechanisms that make them plausible candidates for being natural kinds.

References

Barrett, L. 2017. *How Emotions Are Made*. New York: Houghton Mifflin Harcourt.

---. 2012. "Emotions Are Real." *Emotion* 12: 413-429.

---. 2009. "Variety is the Spice of Life." *Emotion and Cognition* 23: 1284-1306.

---. 2006. "Emotions as Natural Kinds?" *Perspectives on Psychological Science* 1: 28-58.

Ben-Ze'ev, A. 2000. *The Subtlety of Emotions*. Cambridge.

Brader, T. et al. 2008. "What Triggers Public Opposition to Immigration?" *American Journal of Political Science* 52: 959-978.

Ekman, P. & D. Cordaro. 2011. "What is Meant by Calling Emotions Basic." *Emotion Review* 3: 364–370

Kihlstrom, J.F. 1999. "The Psychological Unconscious." In L.A. Pervin & O.P. John (Eds.), *Handbook of Personality* (2nd ed., pp.424–442). New York: Guilford Press.

Kurth, C. 2018. *The Anxious Mind*. MIT Press.

LeDoux, J. 2015. *Anxious*. New York: Viking.

MacKuen, M. et al. 2010. "Civil Engagements," *American Journal of Political Science* 54: 440–458.

Olson, J. 1988. "Missattribution, Preparatory Information, and Speech Anxiety" *Journal of Personality and Social Psychology* 54: 758-767.

Reisenzein, R. 1983. "The Schachter Theory of Emotion" *Psychological Bulletin* 94: 239-264.

Russell, P. 2004. "Core Affect and the Psychological Construction of Emotion. *Psychological Review* 110: 145–172

Scarantino, A & P. Griffiths. 2011. "Don't Give Up on Basic Emotions" Emotion Review 3: 1-11.

Singerman, K. et al. 1976. "Failure of a 'Misattribution Therapy' Manipulation with a Clinically Relevant Target Behavior" *Behavior Therapy* 7: 306-316.

Slivken, K. & A. H. Buss. 1984. "Misattribution and Speech Anxiety" *Journal of Personality and Social Psychology* 47: 396-402.

Valentino, N. et al. 2008. "Is a Worried Citizen a Good Citizen?" *Political Psychology* 29: 247–73.

Winkielman, P. et al. 2005. "Unconscious Affective Reactions to Masked Happy versus Angry Faces Influence Consumption Behavior and Judgments of Value." *Personality and Social Psychology Bulletin* 121-135.

Wong, M. 2017. "The Mood-Emotion Loop" *Philosophical Studies* 173: 3061-3080.

Symposium: *Bridging the Gap Between Scientists and the Public*, PSA 2018

How trustworthy and authoritative is scientific input into public policy deliberations?[i]

Hugh Lacey
Swarthmore College / University of São Paulo

Abstract: Appraising public policies about using technoscientific innovations requires attending to the values reflected in the interests expected to be served by them. It also requires addressing questions about the efficacy of using the innovations, and about whether or not using them may occasion harmful effects (risks); moreover, judgments about these matters should be soundly backed by empirical evidence. Clearly, then, scientists have an important role to play in formulating and appraising these public policies.

However, ethical and social values affect decisions made about the criteria (1) for identifying the range of risks, and of relevant empirical data needed for making judgments about them, that should be considered in public policy deliberations, and (2) for determining how well claims concerning risks should be supported by the available data in order to warrant that they have a decisive role in the deliberations. Consider the case of public policies about using GMOs. Concerning the range of data: is it sufficient for risk assessment only to be informed by data relevant to investigating the risks of using GMOs that may be occasioned by way of physical/chemical/biological mechanisms directly triggered by events within their modified genomes? Or: should data pertaining to the full range of ecological and socioeconomic effects of using them, in the environments in which they are used and under the socioeconomic conditions of their use, also inform this assessment? Those interested in producing and using GMOs, in the light of their adhering to values of capital and the market, are likely to give a positive answer to the first question; those holding competing values, e.g., connected with respect for human rights and environmental sustainability, to the second. And, concerning the degree of support: the former – citing the ethical gravity of losses (both economic and, allegedly, for food security) that would be incurred by failing to use GMOs on a wide scale – are likely to require less stringent standards of evidential appraisal than the latter.

Scientists, *qua* scientists, however, do not have special authority in the realm of values. Thus, their judgments, about the evidential support that claims about risks (and some other matters) have, may sometimes be reasonably (although not decisively) contested partly on value-laden grounds – as they have been in the GMO case, where the contestation has generated considerable controversy, and continues to do so. It follows that, in the context of deliberations about public policy, unless scientists engage with representatives of all stakeholders in the outcomes of the policies (as, for the most part, has not happened in the GMO case) – taking into account that their competing values may lead to making different decisions about what are the relevant data, as well as about the degree of support required for their claims about risks to gain the required credibility to inform the deliberations; and respecting "tempered equality" of participants in the dialogue (Longino) – their trustworthiness is put into question and their authority diminished.

1.

In a letter, dated June 29th, 2016, 135 Nobel laureates made the following claims, among others,[ii] related to using GMOs (genetically modified organisms) in agriculture:

(i) "Scientific and regulatory agencies around the world have repeatedly and consistently found crops and foods improved through biotechnology to be as safe as, if not safer than those derived from any other method of production."

(ii) "There has never been a single confirmed case of a negative health outcome for humans or animals from their consumption."

(iii) "Their environmental impacts have been shown repeatedly to be less damaging to the environment, and a boon to global biodiversity" (Laureates Letter, 2016).

Reflecting the authority and esteem that tends to be accorded to Nobel laureates, the declaration was widely reported and taken to bolster the allegation that there is a *scientific consensus* that cultivating and harvesting genetically engineered crops, and consuming their products, is safe.[iii] The scientists who signed it aimed to assure the public that the three claims are well con-

firmed, and that public policy and regulatory deliberations should reflect them. The claims do not derive from outcomes of the research conducted by these scientists, for at most one or two of them (so far as I can tell, none) have themselves engaged in biosafety research. They were putting their authority behind the research and judgments of others, whom presumably they trusted. Even so, one might reasonable assume that they had, before signing the declaration, examined the relevant research and concurred with its outcomes, and had found good reason to tell us, as they do, (presumably based on a thorough examination of its writings and actions) that the opposition is "based on emotion and dogma contradicted by data" and that it "must be stopped." At the end of the paper, I will argue that the declaration misuses scientific authority and contributes to doubts about the trustworthiness of leading scientific authorities. My larger purpose, however, is to suggest **some** necessary conditions for re-establishing trust in scientific communities – bridging the gap between scientists and the public, and (the concern of de Martín-Melo & Intemann, 2018) – so that both the authority and integrity of science, and the conditions for strengthening democratic societies, are enhanced

2.

First, some more general remarks. I maintain that the deliberations out of which arise public polices having to do with introducing, using and regulating technoscientific innovations (I only have time to discuss GEOs) should consider:

(1) questions about the *efficacy* of the proposed uses are addressed – and about their *safety*, specifically about how well available empirical evidence confirms that the proposed uses do not occasion harmful effects (or risks of causing harmful effects);

(2) the values reflected in the interests expected to be served by the proposed uses, as well as questions about whether interests expected to be served by competing values may be disadvantaged by them, and priorities among the competing interests;

(3) identified potential alternatives to using these innovations – including fundamentally different kinds of practices – as well as how using them compares to the proposed uses with respect to efficacy and safety (and other potential benefits).[iv]

Of these conditions only (1) is uncontroversial and generally followed (although there are disagreements about how it ought to be followed) in public policy deliberations.[v] Clearly satisfactory answers to the questions about efficacy and safety depend on trustworthy and reliable scientific input. I will not question that scientific research has reliably established the efficacy of the GEOs that have already been approved by regulatory bodies for agricultural use, for the most part GEOs with herbicide-resistant and insecticidal properties.[vi] Efficacy does not imply safety, however, and the research approaches (in molecular biology, biotechnology, etc) within which efficacy is established do not suffice for engaging in research dealing with safety. However, many regulatory practices presuppose that scientific input, pertaining to deliberations about safety – like that about efficacy – is obtained prior to consideration of (2) and (3), and to entanglement with value questions. Hence, the currency of the terms "scientific risk assessments" and

"scientific safety studies", areas of research in which scientific/technical "experts" should be granted authority.

One needs to be wary here, for "safe" and "risk" are 'thick ethical terms'. Scientific safety studies cannot be fully separate from entanglement with values and obligations. Thus, e.g. (simplifying a little), 'using X *is* unsafe' implies (*ceteris paribus*) 'X *should* not be used, unless appropriate precautions are taken.' And, when scientists conclude, on the basis of their investigations, that 'using X is safe', they intend it to follow (and to have impact at step (2)), that *ceteris paribus* 'it is improper to impede using X'.[vii] This does not mean that, in the course of empirical research in scientific safety studies, value-laden terms are used in articulating hypotheses and reporting empirical data. The link between the results of the empirical research and the subsequent value judgments depends on a step (call it step (0)), casually made prior to the empirical investigations. At step (0), the set of possible unintended collateral effects of using X is scrutinized, and those that are identified as harmful (as risks)[viii] – obviously value judgments are made here – are then investigated for such matters as the probability and magnitude of their possible occurrence, and its being countered by introducing scientifically informed regulations. In the investigation, the possible collateral effects are characterized, not with thick ethical terms, but with theoretical and observational terms deployed in relevant scientific fields, like molecular biology, chemistry, soil sciences and physiology (whose terms have no value connotations). Then, 'using X is safe' may be concluded,[ix] – usually qualified by 'provided that it is used in accordance with stipulated regulations' – if the investigations confirm that none of the investigated effects would occur with significant magnitude and probability when X is used in accordance with the regulations. This account is consistent with the picture of scientific safety studies that has step (1) preceding steps (2) and (3); but it clarifies that the move from empirically confirmed results at (1) to the claim the value-implicated 'X is safe' and to value judgments of relevance at (2) rests upon value judgments made at step (0). It follows that the conclusion, 'X is safe', might appropriately be challenged – without thereby challenging the scientists' judgments about each of the particular possible effects investigated – on the basis of the value judgment that not all the harmful possible effects of using X were identified at (0).

The outcomes of "scientific" safety studies usually constitute the only input to the deliberations of the 'technical' commissions that participate in public policy deliberations about using and regulating technoscientific objects. In these studies (in the GEO case), at step (0), the possible effects identified as harmful are a subset of those that may be occasioned by way of physical/chemical/biological mechanisms directly triggered by events within the modified genomes of plants. One can identify *two ways in which the adequacy of these studies might be challenged*.[x]

First: Conclusions drawn about the safety of using V (a genetically engineered plant variety) could be challenged on the ground that the subset chosen for investigation does not include some possible effects, with similar mechanisms, that are of of special salience for those who uphold a particular value-outlook.[xi] For them, even well conducted studies on the the items of the subset chosen will be insufficient to confirm that using V is safe.[xii] Challenges of this type can be

resolved (in principle) by conducting more scientific studies of the same kind after having identified a larger relevant subset.[xiii]

Second: Their adequacy could be challenged by those, who object that the set from which the subsets are chosen for "scientific safety studies" is not sufficiently encompassing. For them, deliberations about the safety of using GE-plants should be informed by appropriate empirical investigations, not only of potential effects occasioned by way of physical/chemical/biological mechanisms directly triggered by events within their modified genomes, but also the full range of potential ecological and socioeconomic effects occasioned by using them in the environments (agroecosystems) of their actual or intended use, and under the socioeconomic conditions of their use, taking fully into account that the potential effects vary from variety to variety and species to plant species. Upholding values of respect for human rights, democratic participation and environmental sustainability, which are opposed to those of capital and the market, often motivates challenges of this kind. These potential effects cannot *all* be investigated in "scientific safety studies," for they require utilizing ecological, human and social categories that have no place in research in such areas as physics, chemistry, and molecular biology, and that may include thick ethical terms (e.g., food security, being poisoned).[xiv] To investigate them empirically, therefore, requires adopting methodological approaches that are not reducible to those used in the indicated scientific areas, and that are generally outside of the expertise of scientists trained in the methodologies appropriate to them. The expertise required to engage in research that leads to the development of GEOs is quite different from that required for studies about the safety of using them.

At issue here are not only concerns about risks (potential harmful effects). Farmers (and their communities) in many areas of the world have suffered serious health problems because of having been exposed to glyphosate (the principal active ingredient in the widely used herbicide, RoundUp) sprayed on fields planted with glyphosate-resistant GEOs.[xv] They are unimpressed when told that the varieties of GEOs planted in these fields had undergone and passed "scientific safety tests." They know from their experience (even if it is not well recorded in peer reviewed studies) that, regardless of what was the case in the conditions of the tests, it is not safe to cultivate these GEOs (which require the accompanying use of glyphosate) in the ways and under the conditions in which they are used in their locales. And, they continue to be unimpressed when the manufactures and regulators of the GEOs insist that the problem was not with cultivating the GEOs, but with using glyphosate without heed to stipulated regulations for safe use,[xvi] for they have good reason to believe that the sellers of GEOs and glyphosate know that they will in fact not be used in accordance with these regulations.[xvii]

3.

Summing up, ethical and social values properly affect decisions (at step 0)) made about the criteria to be deployed for identifying the range of risks that should be considered in public policy deliberations, and of the relevant kinds of empirical data needed for making judgments about them. They also – consistent with maintaining that judgments about safety (step (1)) can be settled

prior to steps (2) and (3) – also affect the standards deployed for determining how well claims about risks should be supported (by the available empirical data) – in order to ensure that risks are dealt with properly in public policy deliberations.

Those who uphold values of capital and the market (agribusiness corporations, governments that prioritize economic growth, etc) are likely to cite the ethical gravity of losses (both economic and, allegedly, for food security) that would be incurred by failing to use GEOs on a wide scale; and consequently to require less stringent standards of evidential appraisal than those who uphold values of respect for human rights, democratic participation and environmental sustainability, who are likely to adopt precautionary stances that permit time for research incorporating more stringent standards to be met.[xviii] Similarly, those who uphold the latter values are likely to emphasize the importance of step (3): investigating alternatives to the food/agricultural system, in which using GEOs and the use of agrotoxics are acquiring ever larger roles, alternatives such as agroecology, a scientifically-informed approach to agriculture that attends simultaneously to production, sustainability, social health, strengthening the values and cultures of local communities, and to furthering the practices needed to implement policies of food sovereignty – and to urge the public support of research, in which are adopted strategies appropriate for dealing with the human, ecological and social dimensions of agroecosystems.[xix]

Scientists, *qua* scientists, however, do not have authority in the realm of ethical and social values. The values they uphold, even when widely shared, do not trump those upheld by other groups in democratic public policy deliberations. Thus, their judgments, about the evidential support that claims about the safety of planting GEO crops and consuming their products have, may sometimes be reasonably contested partly on value-laden grounds (cf. de Melo-Martín & Intemann, 2017, p. 131). That contestation cannot be rebutted by appeal to the alleged "scientific consensus" that GEOs (or, particular varieties of them) are safe. Apart from the fact that actually there is no such consensus, manifestly so among experts in biosafety investigations,[xx] if there were, it would likely secrete the scientists' shared value commitments, a matter on which they have no authority. Appeal to such an alleged consensus covers up the role of upholding the values of capital and the market in affirming it.

It follows that, in the context of deliberations about public policy, the trustworthiness of scientists is put into question and their authority unmerited,

- unless they engage with representatives of all stakeholders in the outcomes of the policies (as, for the most part, has not happened in the GEO case);
- unless, moreover, in doing so – respecting what Longino (2002, p. 129–135) calls "tempered equality" of participants in the deliberations – , they take into account that upholding competing values (e.g., of company-employed scientists and family farmers) may lead to making different judgments concerning relevant data, hypotheses to investigate, and approaches to farming, as well as concerning the degree of support required for claims about safety to merit credibility.

4.

Let us now return to the three claims (introduced at the outset) that the 135 Nobel laureates endorsed:[xxi]

These claims are ambiguous, misleading, in some instances false, and apparently made without acquaintance with the relevant studies and arguments of their critics. (i) is false: I am not aware of any agency that has compared the safety of GEO crops and their food products with that of agroecological (or organic farming) methods of production – the agencies have not sought out the results of research dealing with that comparison (and very little of it has been conducted). At most, they have found GEO crops and products to be at least as safe as conventional high-input crops and their products, but that doesn't respond to the critics who endorse agroecological methods of production. (ii) is probably true – but misleading: it does not mention that epidemiological studies of consumption of GEOs have not been conducted,[xxii] to a large extent because legal prohibition of labelling GEO products poses probably an insurmountable impediment to conducting them; and that it is well documented that cultivating GEOs has occasioned health problems for numerous farmers who have been exposed to the agrotoxics, whose use is integral to the cultivation of certain varieties of GEOs. (iii) is ambiguous: the environmental impacts may indeed be less damaging than those of conventional high-input agriculture; but they are incomparably more damaging to the environment than agroecological farming that has environmental sustainability built into its fundamental objectives.

By dismissing criticisms like these "based on emotion and dogma contradicted by data," and not attempting to rebut them in a context where something like Longino's conditions are in place, the scientists undermine the authority that science should be able to demand to be recognized; and they weaken the contribution that science could make to democratic policy deliberations.

References

Bombardi, Larissa M. (2017) Geografia do Uso de Agrotóxicos no Brasil e Conexões com a União Europeia. E-book, <https://drive.google.com/file/d/1ci7nzJPm_J6XYNkdv_rt-nbFmOETH80G/view>. São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

De Melo-Martín, I. and Intemann, K. (2018) *The Fight against Doubt: How to bridge the gap between scientists and the Public*. New York: Oxford University Press.

Hilbeck, A., Binimelis, R., Defarge, N., Steinbrecher, R., Székács, A., Wickson, F., Antoniou, M., Bereano, P. L., Clark, E. A., Hansen, M., Novotny, E., Heinemann, J., Meyer, H., Shiva, V. & Wynne, B. (2015) No scientific consensus on GEO safety. *Environmental Sciences Europe* 27: 4–9.

Human Rights Watch (2018) The Failing Response to pesticide Drift in Brazil's Rural Communities, July 20, 2018, <https://www.hrw.org/report/2018/07/20/you-dont-want-breathe-poison-anymore/failing-response-pesticide-drift-brazils>.

Krimsky, S. (2015) An illusory consensus behind GEO health assessment. *Science, Technology and Human Values* 40 (6): 883–914.

Lacey, H. (2005) *Values and Objectivity in Science; current controversy about transgenic crops.* Lanham, MD: Lexington Books.

— (2015a) Food and agricultural systems for the future: science, emancipation and human flourishing. *Journal of Critical Realism* 14 (3), 2015: 272–286.

— (2015b) Agroécologie : la science et les valeurs de la justice sociale, de la démocratie et de la durabilité. *Ecologie et Politique*, No. 51, 2015: 27–40.

— (2016) Science, respect for nature, and human well-being: democratic values and the responsibilities of scientists today. *Foundations of Science* 21(1): 883–914.

— (2017) The safety of using genetically engineered organism: empirical evidence and value judgments. *Public Affairs Quarterly* 31 (4): 259–279.

Lacey, H., Corrêa Leite, J., Oliveira, M.B., & Mariconda, P.r. (2015a) Transgênicos: malefícios, invasões e diálogo. *JC Notícias*, Edition 5167 (April 30, /2015), http://www.jornaldaciencia.org.br/edicoes/?url=http://jcnoticias.jornaldaciencia.org.br/9-transgenicos-maleficios-invasoes-e-dialogo/.

— (2015b) Transgênicos: diálogo. *JC Notícias*, Edition 5182 (May 22/2015), http://www.jornaldaciencia.org.br/edicoes/?url=http://jcnoticias.jornaldaciencia.org.br/27-transgenicos-dialogo/.

Longino, H. (2002) *The Fate of Knowledge*. Princeton: Princeton University Press.

Laureates Letter (2016) "Laureates letter supporting precision agriculture," http://supportprecisionagriculture.org/nobel-laureate-gmo-letter_rjr.html .

US National Academies of Science, Engineering and Medicine (2017). *Genetically Engineered Crops: Experiences and Prospects*. Washington: National Academies Press.

Paganelli, A., Gnazzo, V, Acosta, H., López, S.L. & Carrasco, A.E. (2010) 'Glyphosate-based herbicides produce teratogenic effects on vertebrates by impairing retinoic acid signaling'. *Chemical Research in Toxicology* 23: 1586–1595.

Traavik, T. & Ching, L.L. (2007) *Biosafety first: Holistic approaches to risk and uncertainty in genetic engineering and genetically modified organisms*. Trondheim, Norway: Tapir Academic Press.

Appendix

The central concern of the letter signed by the Nobel laureates is to support the program of research on Golden Rice [a variety of genetically engineered rice] and to denounce opposition to it, especially that of the NGO, Greenpeace. In a longer work, I would also discuss critically the way in which the letter misleads both about the state of research on Golden Rice and about that character of criticisms that question the importance of this research.

(a) The letter states that Greenpeace "has spearheaded opposition to Golden Rice, which has the potential to reduce or eliminate much of the death and disease caused by a vitamin A deficiency, which has the greatest impact on the poorest people in Africa and Southeast Asia". It called upon "governments of the world to reject Greenpeace's campaign against Golden Rice specifically, and crops and foods improved through biotechnology in general; and to do everything in their power to oppose Greenpeace's actions and accelerate the access of farmers to all the tools of modern biology, especially seeds improved through biotechnology"; and concluded with the warning: "Opposition based on emotion and dogma contradicted by data must be stopped," accompanied by the rhetorical question: "How many poor people in the world must die before we consider this a 'crime against humanity'?"

(b) Around the same time, the US National Academies of Science, Engineering and Medicine (2017) pointed out that the International Rice Research Institute (IRRI) had stated reported: "Golden Rice will only be made available broadly to farmers and consumers if it is successfully developed into rice varieties suitable for Asia, approved by national regulators, and shown to improve vitamin A status in community conditions. If Golden Rice is found to be safe and efficacious, a sustainable delivery program will ensure that Golden Rice is acceptable and accessible to those most in need" (p. 228). As of July 2016, IRRI was continuing research on developing varieties of Golden Rice for use in SE Asia, and (according to it) none of the conditions it stated had yet been met - it is for this reason that Golden Rice has not been introduced.

(c) Two years later, earlier this year (2018), IRRI asked the USFDA for an opinion regarding the safety of a variety of Golden Rice (called GR2E - the only variety yet submitted for regulatory approval - but not yet approved in any Asian country). FDA (May 24, 2018) endorsed the evaluation of IRRI (and the Australian regulatory body) that GR2E is safe for consumption, while pointing out that it is not intended for food or animal uses in USA. However, it added: "the concentration Beta-carotene in GR2E rice is too low to warrant a nutrient content claim." GR2E is safe but not nutritionally relevant.

(d) The signers of the letter, thus, were remarkably uninformed about the state of research on Golden Rice – and also about the views and stances of Greenpeace (I am not associated with Greenpeace). On its website Greenpeace states that its objective is to "ensure the ability of Earth to nurture life in all its diversity." It fits into the body of critics of using GMOs, who maintain that the dominant food-agricultural system (in which using GEOs has become for the time being a fundamental component) cannot respond adequately to the food and nutrition needs of the world's impoverished peoples (and the right to food security for everyone), and that these needs can best be ameliorated by the programs of agroecology and food sovereignty (Lacey, 2015a; 2015b) – and that programs for developing GEOs (like Golden Rice) are taking resources away from developing effective and lasting solutions to death and disease caused by vitamin A deficiency. Greenpeace has a respected place among these critics (and its "direct actions" and contributions to legal challenges are often appreciated by them). Of course, it would be legitimate to rebut the critics with argument and evidence. One wonders why the laureates did not attempt to do so.

(e) The credibility of pronouncements made by scientists of outstanding achievement is weakened when they sign letters like this one, accompanied by inflated, emotionally charged rhetoric, that has has a slender basis in fact. It would be enhanced if they entered into the type of dialogue, advocated by Helen Longino, in which scientists would "listen to" the evidence provided by relevant parties, attempt to understand critics, and not tar them without a hearing. Science has an indispensable contribution to make in policy deliberations; but it is not the determiner of policy. Science will be enhanced, and its role in democratic societies consolidated, if it claims only to have authority where it is actually warranted.

Notes

i **DRAFT** (not for citation outside of the PSA meeting in Seattle) – October 15, 2018. The text is a draft of the presentation I'm planning to make. The notes contain details that will be incorporated into an eventual completed paper.

ii See Appendix.

iii E.g., Mark Lynas (Cornell Alliance for Science), *A plea to Greenpeace*, <http://www.marklynas.org/2016/06/a-plea-to-greenpeace/>.
In this paper I only consider GEOs used in agriculture. I take for granted that claims to the effect that using GEOs is safe refer to GEOs that have passed safety tests, including those currently available on the market. (Obviously an unsafe GEO could be developed. Some varieties of GEOs have been developed that, after failing to pass safety tests, were not released for use.)

iv More fully developed and defended in Lacey (2005), Part 2.

v Deliberations concerning (2) and (3) cannot be settled in scientific inquiry (sound empirical inquiry), but there are sound empirically-based inputs that are (or could be) relevant to them. The deliberations will not be satisfactory if they do not draw upon these inputs. (See Lacey, 2005.)

vi Claims about efficacy need to be stated in a more qualified and nuanced way. I also will not contest that the claim that scientific research has not provided compelling evidence that consuming GEO products is unsafe health-wise. (The absence of compelling evidence that GEO products are unsafe to consume does not mean that there is compelling evidence that they are safe to consume – it depends on whether or not the necessary research has been conducted.)

vii The *ceteris paribus* qualification is needed to take into account that sometimes considerations, not reducible to safety ones, may properly be appealed to.

viii I will not discuss here how this set is generated – e.g., from considering past investigations, role of values in it, stakeholders' concerns, etc) – and who (holding what values?) makes (and should make) the identification of what should be considered harmful? following what kinds of deliberations? and who should be represented in the deliberations?.

ix To conclude on the basis of empirical investigation that 'X is safe' requires showing one-by-one that each member of the set of anticipated effect (judged to be harmful) is unlikely to occur at sufficient magnitude under the conditions imposed by proposed regulations. This presupposes: (a) an inductive move to unanticipated effects; and (b) that representative cases of all the effects, that should be labelled potentially harmful, are members of the set.

x I have argued elsewhere that here methodological and value considerations mutually reinforce each other (Lacey, 2017). Proponents of using GEOs often say that these safety studies investigate the risks occasioned by the GEOs themselves, and not those occasioned by the accompaniments of using them in agroecosystems or by socioeconomic mechanisms.

xi E.g., effects on soil microorganisms, a matter especially salient for those who regard maintaining soil fertility as indispensable for sustainable agriculture.

xii The studies, which have produced many of the results that have actually informed public policy and regulatory decisions, have been criticized for having a number of kinds of shortcomings (e.g., connected with conflicts of interest, and the use of intellectual property rights to maintain studies secret and so unavailable for replication and independent confirmation). Value judgments pervade these criticisms and their rebuttals. I will not attend to the questions that arise here.

xiii Such challenges might be deemed irrelevant by those who reject the value-outlook for which the possible effects have special salience, and so who reject the need for the further studies. Those adhering to the values of capital and the market sometimes take such a stand. How reasonable that might be depends on the arguments offered against holding the value-outlook in question.

xiv For elaboration see Lacey (2016; 2017).

xv For documentation, see, e.g., Bombardi (2017); Paganelli, et al. (2010); Human Rights Watch (2018).

xvi After a jury in California recently ruled that Monsanto was responsible for a man's being afflicted with cancer, and imposed a huge fine on it because it – for it was deemed that Monsanto had "acted with malice" in not providing warning on its label of the risks to health occasioned by using Roundup – the President of Bayer (that has now incorporated Monsanto) responded: "The correct use of Roundup doesn't present a risk to health" (reference to be added). [Monsanto has appealed the ruling.]

xvii Three years ago, when representatives of farmers – who had been poisoned in this way – came to present their testimony at a meeting of the "technical" commission in Brazil (CTNBio) that had appraised a particular variety of GEOs as safe, they were not granted a hearing since (most members of the commission maintained) they were bearers only of anecdotal (not scientific) evidence that had no relevance to the conclusions of scientific safety studies. When they then disrupted the meeting (and others of their group prevented the planting of a new variety of GEOs by invading a nursery and pulling up all the seedlings), they were denounced by major scientific organizations as having no respect for science, and acting on the

basis of "emotion and dogma." For criticisms of this stance taken by the majority of members of CTNBio, and a response to a rebuttal of the criticism, see Lacey, et al. (2015a; 2015b), articles published in *JC Notícias*, a daily e-newsletter of *Jornal da Ciência*, a publication of SBPC (Brazilian Society for the Advancement of Science).

The narrow scope of "scientific safety studies" is sometimes justified on the ground that the investigations of the social impact of using GEOs is not "scientific," for the methodologies adopted in them are not reducible to those adopted in the mainstream areas of science mentioned above. Be that as it may: I won't quibble about how to use the term "scientific" (a thick ethical term); the investigations in question are (when properly conducted) systematic empirical investigations. If they don't count as "scientific", that would imply that the results of "scientific" investigations cannot provide sufficient input into deliberations concerning public policies about safety, and would need to be supplemented with input from other kinds of empirical investigations.

[xviii] See Lacey (2017).

[xix] For details, see Lacey (2005; 2015a; 2015b).

[xx] See, e.g., Hilbeck, et al. (2015); Krimsky (2015); Traavik & Ching (2007).

[xxi] See Appendix.

[xxii] Unless all the relevant research has been conducted (and it has not been in this case), the absence of compelling evidence that GEO products are unsafe to consume does not imply that there is compelling evidence that they are safe to consume – and it has nothing to do with harms that may be caused by, e.g., contact with an agrotoxic, rather than by consumption.

The Reference Class Problem for Credit Valuation in Science

Carole J. Lee (c3@uw.edu)

Abstract: Scholars belong to multiple communities of credit simultaneously. When these communities disagree about how much credit to assign to a scholarly achievement, this raises a puzzle for decision theory models of credit-seeking in science. The reference class problem for credit valuation in science is the problem of determining to which of an agent's communities – which reference class – credit determinations should be indexed for any given act under any given state of nature. I will identify strategies and desiderata for resolving ambiguity in credit valuation due to this problem and explain how pursuing its solution could, ironically, lead to its dissolution.

1. Introduction

Within the scientific community, there is a common understanding that its reward system drives problematic behavior linked to publication patterns, pipeline retention, hypercompetitive scientific cultures, and reproducibility. Conversely, there is also a shared sentiment that, in order to change these cultures and behaviors in ways that would improve science, the scientific community must coordinate across institutions to change how credit is assigned at the level of the individual scientist (Alberts et al. 2014, Nosek et al. 2015, Aalbersberg et al. 2017, National Academies of Sciences 2018, National Science Foundation 2015, Blank et al. 2017). The hope is

that increasing individual researchers' incentives towards increased transparency and openness will improve the integrity, reproducibility, and accuracy of the published record.[1]

Analogously, philosophers working in the "credit economy" tradition adopt the working assumption that there is some amount of credit that agents can accrue for different acts under different states of nature. This assumption allows them to use decision theory to model how credit-seeking among individual scientists can give rise to behavior and norms that support or thwart the achievement of community-wide goals. When, in the aggregate, individual credit-seeking cuts against collective ends, their approach can explore how changes to individuals' incentive structures can nudge and redirect individual behavior (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Bright 2017, Heesen 2017, Kitcher 1990, Strevens 2003, Zollman 2018). Different philosophers make different assumptions about the norms by which credit gets allotted – for example, whether credit is best thought of as all-or-nothing (Strevens 2003, Bright 2017, Heesen 2017) or as something that may come in degrees (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Zollman 2018). However, the general approach assumes that there is some precise way to assign credit to different acts under different states of nature – an assumption that allows these philosophers to model credit-seeking behavior and the emergence of scientific norms in formally tractable ways.

But, how much credit gets assigned to any given act under any given state of nature? Just as each of us simultaneously belongs to multiple social categories each of which is tied to implied social hierarchies (Macrae, Bodenhausen, and Milne 1995, Crenshaw 1989), each

[1] Institutions can also experience incentives that promote or thwart scientific ends (Lee and Moher 2017).

scholar simultaneously belongs to multiple communities of value with implied social hierarchies for assigning credit. To which of an agent's communities – which reference class – should credit determinations be indexed and why?

In this paper, I will use examples from the current context of science's complex and dynamic culture to motivate and illuminate what I will call the *reference class problem for credit valuation in science*. I will identify a few strategies and desiderata for solving ambiguity in credit assignments due to the reference class problem. And, I will say a bit about how developing the resources needed to solve it could ultimately sow the seeds for its own dissolution.

2. The Reference Class Problem for Credit Valuation in Science

The contours of this puzzle about the "coin of recognition" (Merton 1968, 56) become visible when one moves beyond thinking about credit in generic, abstractions of scientific communities towards the heterogeneous communities we find today. I start from this slightly more concrete perspective because prestige requires recognition *by individuals and forums* that are themselves valued by credit-seeking scholars (Zuckerman and Merton 1971, Lee 2013): credit worthiness in science is a function of the individuals and systems designed to assess, allocate, dispute, and enforce it. Although some aspects of Zuckerman and Merton's narrative about the origins of the normative structure of science have been contested by historians (Csiszar 2015, Biagioli 2002), we see the social dynamics Zuckerman and Merton proposed clearly at play in contemporary science. For example, Nature Publishing Group recently found that – for the 18,354 authors in science, engineering, and medicine surveyed – the reputation of a journal is the primary factor driving choices about where to submit their work, where reputation is

primarily determined by the journal's impact factor and whether it is "seen as the place to publish the best research" (Nature Publishing Group 2015). Factors associated with a journal's ability to archive and disseminate research – things like a journal's time from acceptance to publication, indexing services, or Open Access options – were much less important.[2]

Within academia, each of us simultaneously belongs to multiple communities of value. The reference class problem arises when these different communities of value disagree about the amount of credit an agent accrues for choosing some act under some state of nature. Although I take this problem is be general, for the sake of clarity and simplicity in presentation, I will focus my examples on communities that can be described as having a nesting structure: for example, individual scholars belong to specific sub-disciplines, which are nested within disciplines, which are nested within a more general population of scholars. A sub-population that is nested within a population can have a credit sub-culture whose valuations differ from that of the population, whose valuations can differ from that of the super-population. In these cases, changing how narrowly or broadly one draws the boundaries of an agent's community of valuation can change the amount of credit assigned to a scholarly accomplishment. This gives rise to the *reference class problem for credit valuation in science*: to which of the agent's communities – which reference class – should credit valuations be indexed when determining the amount of credit the agent accrues for different acts under different states of nature?

[2] I recognize that some decision theorists, especially those working outside of philosophy, may reject or remain agnostic about attributing mental states such as beliefs to agents (Okasha 2016). However, because I understand credit and credit-seeking as sociological phenomena involving status beliefs such as these, I am committed to attributing beliefs to agents.

There are many examples across academia where nesting community structures can give rise to paradoxes and pathologies in credit assignments. For example, scholars' individual sense of what counts as quality work – their individual credit assignments – may deviate from what is endorsed in a sub-discipline or discipline's status hierarchy (Correll et al. 2017, Centola, Willer, and Macy 2005, Willer, Kuwabara, and Macy 2009). A puzzle that has cachet in a sub-discipline may be of peripheral importance within that discipline: for example, a more accurate technique for measuring how temperature cools with elevation considered critical in mountain meteorology and mountain ecology (Mindner, Mote, and Lundquist 2010) may have less visibility, despite its relevance, to the larger discipline of hydrology (Livneh et al. 2013). A question or technique that is thought to have high impact across fields (e.g., machine learning) may have little prominence within some of those fields.

Hypothetically speaking, one could imagine differences in valuations giving rise to a *Simpson's paradox in credit valuation*. Simpson's paradox is a phenomenon whereby a trend that appears in a population reverses or disappears when it is disaggregated into sub-populations (Blyth 1972). For example, a classic study found that, when looking at aggregate graduate school admissions data at UC Berkeley, women were, on the whole, less likely than men to be accepted; however, when the data was disaggregated into admitting departments, women were more likely than men to be admitted (Bickel, Hammel, and O'Connell 1975). Analogously, a *Simpson's paradox in credit valuation in science* would occur in cases where a population-level preference for scholarly product *a* versus *b* reverses when the population is disaggregated into its component sub-populations. In Simpson's Paradox cases, thinking more carefully about the context of evaluation usually leads to using a reference class that is finer-grained than the population-level. However, it's not clear whether this would always be the case in evaluations of

scientific credit. Hypothetically speaking, consider a hypothetical scenario in which an interdisciplinary project is not preferred by the individual disciplines represented by its authors or content, but is preferred when those disciplines are aggregated together. And, imagine that this project gets published in a journal, valued by those disciplines, that seeks papers of interest *across and beyond disciplines* (not just within disciplines): this is one way to interpret, for example, *Science's* mission to publish papers that "merit recognition by the wider scientific community and general public. . . beyond that provided by specialty journals" (Science). Which reference class would be most relevant in evaluating the value of this project?

There are other ways of dividing scholarly communities into nesting structures that create tensions in credit assignments. The pressures a scholar may feel from the incentive structure impacting her department/school may be slightly different from the incentive structure impacting her university. A coarse but concrete way to see this is to think about the prestige structure reified and reinforced by ranking systems (Espeland and Sauder 2012, 2016, Sauder and Espeland 2006), which transform "the ways professional opportunities are distributed" (Espeland and Sauder 2016, 7). An untenured business school professor with a potentially high impact manuscript needs to burnish her prestige in the eyes of both her dean and her provost, since both will evaluate her tenure case. If her provost is working to gain stature on the Academic Rankings of World Universities [ARWU], the professor should submit her manuscript to *Science* or *Nature*, since the ARWU ranks universities by their publications in these journals (Academic Ranking of World Universities 2018). However, if her dean is trying to gain stature on the *Financial Times* International ranking of MBA programs, she should submit to one of the fifty business, economics, or psychology journals by which the FT ranking system evaluates Business

school prestige – notably, the journal list does not include *Science* or *Nature* (Ormans 2016). What should the business school professor do?

Finally, credit assignments can vary depending on how long a time window a scholar keeps in view. A coarse but concrete way to think about this is by looking at how metrics for evaluating scholarship change over time. Journal impact factors are becoming less useful measures for evaluating an individual's scholarly contribution: since the advent of the digital age, the most elite journals (including *Science* and *Nature*) are publishing a decreasing percentage of the top cited papers (Larivière, Lozano, and Gingras 2013); the relationship between journal impact factor and paper citations has declined over time (Lozano, Larivière, and Gingras 2012); and, the citation distributions between journals "overlap extensively" (Larivière et al. 2016). The current wisdom is that if quantitative indicators are to be used to evaluate research, it is more useful to use article-level metrics such as citations as well as alternative metrics such as downloads and views (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017). On the horizon, there are now calls for creating new metrics that can encourage researchers and journals to be transparent and open in their reporting practices (National Academies of Sciences 2018, Wilsdon et al. 2017, Aalbersberg et al. 2017). Note that, the rise of such metrics – as well as the growing meta-research literature that ranks journals by the replicability (Schimmack 2015) or sample size and statistical power of their published results (Fraley and Vazire 2014) – makes it possible for a journal's impact factor and epistemic credibility to come apart (Fang and Casadevall 2011).

Decision theorists capture the risky nature of individual choices by allowing for uncertainty about which states of the world will come to be; and, when the probabilities attached to different outcomes are understood subjectively, these models permit a kind of subjectivity in

estimates of expected credit for different acts. However, I hope the examples throughout this section animate genuine *ambiguity in credit* due to the reference class problem for credit valuation in science.

3. Strategies and Desiderata for Solving the Reference Class Problem

How might decision theorists try to solve the reference class problem for assigning credit in science? One possible approach argues for the "correctness" of using one community rather than another. For example, it might be tempting to argue that all prestige is discipline-based since many scholarly prizes are distributed for excellence in particular disciplines (e.g., Nobel prize, Fields prize, academic society prizes); and, even when research is funded or published in interdisciplinary contexts, it may be primarily evaluated on the basis of its disciplinary excellence (Lamont 2009, but see Lee et al. 2013). Indexing credit valuation to a particular community need not prevent scholars from outside that community from understanding the relative value of that contribution: for example, if one were to adopt the old-fashioned and problematic assumption that an article's impact can be measured by the impact factor of the journal in which it is published,[3] and one recognizes that citations rates vary across disciplines, one could use field-normalized percentiles to understand a paper's impact in a metric that is legible across fields (Hicks and Wouters 2015). Because this strategy for addressing the

[3] The citation distributions within journals are so skewed that it is statistically improper to infer the impact of an individual article on the basis of the impact factor of the journal in which it is published (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017, Larivière et al. 2016, Wilsdon et al. 2015).

reference class problem relies heavily on identifying the "right" community, defending the centrality of the chosen community as opposed to others is critical. For example, some may challenge the idea that disciplines should be the sole arbiter of credit: note that the awarding of some scientific prizes reach across disciplinary conceptions of excellence (e.g., consider winners of the MacArthur Genius Prize and the psychologists who have won the Nobel Prize in Economics).

Another possible approach creates an algorithm that calculates the credit value of a scholarly contribution by summing the credit valuation of multiple communities. This approach would need to identify exactly how much to weight each community's valuation – with a rationale for why – since different weightings could lead to different overall credit valuations.[4] Note that some scholars take this style of approach when trying to measure the relative prestige of journals: in particular, the Eigenfactor score rates journals according to the number of its incoming citations, where the "relative importance" of each incoming citation is contextualized by the frequency with which the citing journal is itself cited (West, Bergstrom, and Bergstrom 2010).

Those who may wish to model the implications of different approaches for solving the reference class problem may try to do so by setting up hypothetical communities that assign

[4] On the face of it, this may seem like a form of commensuration because it involves summing values to calculate an overall score (Espeland and Stevens 1998). However, the process of commensuration requires combining values across *qualitatively* different domains of value. For clearer examples of commensuration in scholarly evaluation, see Lee (2015).

community boundaries and credit assignments in *de facto* ways to see what kinds of behaviors and norms emerge.

However, to solve the underlying conceptual problem, one must provide theories of community and credit that address two fundamental but vexing questions. How should one define and gerrymander the boundaries of the relevant communities invoked in the proposed solution? And, how does one determine the amount of credit those communities would assign to different acts under different states of nature? These questions may not be independently answerable. The boundaries of a community may need to be defined in terms of patterns of shared lore among its members about how credit is accrued – shared beliefs that coordinate credit-seeking and enforcement behavior in cases where status beliefs are internalized as norms (Merton 1973) and in cases where they are not (Willer, Kuwabara, and Macy 2009, Ridgeway and Correll 2006). Conversely, in recognition that some community members can have more influence than others on the content of reigning status beliefs, a community's credit assignments may need to be defined with some reference to the causal patterns of interaction among specific individuals and clusters of individuals – including status judges who wield "social control through their evaluation of role-performance and their allocation of rewards for that performance" (Zuckerman and Merton 1971, 66). Note, however, that answers to these questions should not *exclusively* inform each other. Notably, we must be careful not allow the size of a scholarly population and/or the power of its status judges to fully determine the intellectual value of the questions pursued by any particular partition of the scholarly universe.

4. Conclusion

Scientific credit – the "coin of recognition" (Merton 1968, 56) – is assessed, allocated, disputed, and enforced by many different communities and institutions within science that support and sustain a multiplicity of status hierarchies. This gives rise to what I have called the reference class problem for credit valuation in science. Solving this problem requires developing rich theories of community and credit that are based on fine-grained information about the structure and status systems of complex scholarly networks. The irony of this assessment is that such investigation towards solving the reference class problem could ultimately sow the seeds for its own dissolution.

In particular, such study can render friable a critical assumption for both the reference class problem and for decision theory models: namely, that communities, once defined, assign determinate amounts of monistic credit for different acts under different states of nature – that credit "can vary quantitatively but not qualitatively" (Anderson 1993, xii).[5] Contrary to this, recent policy papers call for moving away from narrowly conceived measurements of research excellence towards broader ones that are sensitive to the diversity of individual researchers', programs', and academic institutions' research missions (Hicks and Wouters 2015, Wilsdon et al. 2015). Such work can include community-engaged scholarship that creates, disseminates, and implements knowledge in coordination with the public to identify social interventions, change social practice, and influence policy (Hicks and Wouters 2015, San Francisco Declaration on Research Assessment 2013, Boyer 1990, Escrigas et al. 2014). From the

[5] Note too that, for formal reasons, the assumption that individual credit assessments could be aggregated into a collective one is questionable given the challenges of combining individual preferences into collective ones (Arrow 1950).

perspective of these efforts, plurality in our notions of scholarly excellence and credit – and differences in valuation and prioritization practices between individuals and communities – may be best conceived, not as a logical problem to solve, but as a starting point for theorizing.

Acknowledgments: Many thanks to Christopher Adolph, Aileen Fyfe, Crystal Hall, Jessica Lundquist, Conor Mayo-Wilson, and Kevin Zollman for helpful conversations. This research used statistical consulting resources provided by the Center for Statistics and the Social Sciences, University of Washington.

References

Aalbersberg, IJsbrand Jan, Tom Appleyard, Sarah Brookhart, Todd Carpenter, Michael Clarke, Stephen Curry, Josh Dahl, Alex DeHaven, Eric Eich, Maryrose Franko, Len Freedman, Chris Graf, Sean Grant, Brooks Hanson, Heather Joseph, Véronique Kiermer, Bianca Kramer, Alan Kraut, Roshan Kumar Karn, Carole Lee, Aki MacFarlane, Maryann Martone, Evan Mayo-Wilson, Marcia McNutt, Meredith McPhail, David Mellor, David Moher, Alison Mudditt Mudditt, Brian Nosek, Belinda Orland, Tim Parker, Mark Parsons, Mark Patterson, Solange Santos, Carolyn Shore, Dan Simons, Bobbie Spellman, Jeff Spies, Matt Spitzer, Victoria Stodden, Sowmya Swaminathan, Deborah Sweet, Anne Tsui, and Simine Vazire. 2017. "Making science transparent by default; Introducing the TOP Statement." *OSF Preprints*. doi: https://doi.org/10.31219/osf.io/sm78t.

Academic Ranking of World Universities. 2018. "ShanghaiRanking's Academic Ranking of World Universities 2018 Press Release." accessed September 1.

http://www.shanghairanking.com/Academic-Ranking-of-World-Universities-2018-Press-Release.html.

Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. 2014. "Rescuing US biomedical research from its systematic flaws." *Proceedings of the National Academy of Sciences* 111 (16):5773-7.

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Arrow, Kenneth J. 1950. "A difficulty in the concept of social welfare." *Journal of Political Economy* 58 (4):328-46.

Biagioli, Mario. 2002. "From Book Censorship to Academic Peer Review." *Emergences: Journal for the Study of Media & Composite Cultures* 12 (1):11-45.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. "Sex bias in graduate admissions: Data from Berkeley." *Science* 187 (4175):398-404.

Blank, Rebecca, Ronald J. Daniels, Gary Gilliland, Amy Gutmann, Samuel Hawgood, Freeman A. Hrabowski, Martha E. Pollack, Vincent Price, L. Rafael Reif, and Mark S. Schlissel. 2017. "A new data effort to inform career choices in biomedicine." *Science* 358 (6369):1388-9.

Blyth, Colin R. 1972. "On Simpson's Paradox and the sure-thing principle." *Journal of the American Statistical Association* 67 (338):364-66.

Boyer, Ernest L. 1990. *Scholarship Reconsidered*. San Francisco, CA: The Carnegie Foundation for the Advancement of Teaching.

Bright, Liam Kofi. 2017. "On Fraud." *Philosophical Studies* 174:291-310.

Bruner, Justin, and Cailin O'Connor. 2017. "Power, Bargaining, and Collaboration." In *Scientific Collaboration and Collective Knowledge*, edited by Thomas Boyer-Kassem, Conor Mayo-Wilson and Michael Weisberg, 135-157. Oxford, UK: Oxford University Press.

Centola, Damon, Robb Willer, and Michael Macy. 2005. "The emporer's dilemma: A computational model of self-enforcing norms." *American Journal of Sociology* 110 (4):1009-40.

Correll, Shelley J., Cecilia L. Ridgeway, Ezra W. Zuckerman, Sharon Jank, Sara Jordan-Bloch, and Sandra Nakagawa. 2017. "It's the conventional thought that counts: How third-order inference produces status advantage." *American Sociological Review* 82 (2):297-327.

Crenshaw, Kimberle. 1989. "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics." *University of Chicago Legal Forum* 139:139-168.

Csiszar, Alex. 2015. "Objectivities in Print." In *Objectivity in Science: New Perspectives from Science and Technology Studies*, edited by Flavia Padovani, Alan Richardson and Jonathan Y. Tsou, 145-69. Cham, Switzerland: Springer International Publishing.

Escrigas, Cristina, Jesús Granados Sánchez, Budd Hall, and Rajesh Tandon. 2014. "Editor's introduction. Knowledge, engagement and higher education: Contributing to social change." In *Report: Higher Education in the World*, edited by Cristina Escrigas, Jesús Granados Sánchez, Budd Hall and Rajesh Tandon. Palgrave Macmillan.

Espeland, Wendy Nelson, and Michael Sauder. 2012. "The Dynamism of Indicators." In *Governance by Indicators: Global Power through Quantification and Rankings*, edited by Kevin Davis, Angelina Fisher, Benedict Kingsbury and Sally Engle Merry, 86-109. Oxford: Oxford University Press.

Espeland, Wendy Nelson, and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York, NY: Russell Sage Foundation.

Espeland, Wendy Nelson, and Mitchell L. Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24:313-43.

Fang, Ferric C., and Arturo Casadevall. 2011. "Retracted Science and the Retraction Index." *Infection and Immunity* 79 (10):3855-9.

Fraley, R. Chris, and Simine Vazire. 2014. "The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power." *PLOS ONE* 9 (10):e109019. doi: 10.1371/journal.pone.0109019.

Heesen, Remco. 2017. "Communism and the Incentive to Share in Science." *Philosophy of Science* 84:698-716.

Hicks, Diana, and Paul Wouters. 2015. "The Leiden manifesto for research metrics." *Nature* 520:429-31.

Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* LXXXVII (1):5-22.

Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Larivière, Vincent, Véronique Kiermar, Catriona J. MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. 2016. "A simple proposal for the publication of journal citation distributions." *BioRxiv*:062109.

Larivière, Vincent, George A. Lozano, and Yves Gingras. 2013. "Are elite journals declining?" *Journal of the Association for Information Science and Technology* 65 (4):649-55.

Lee, Carole J. 2013. "The limited effectiveness of prestige as an intervention on the health of medical journal publications." *Episteme* 10 (4):387-402.

Lee, Carole J. 2015. "Commensuration bias in peer review." *Philosophy of Science* 82:1272-83.

Lee, Carole J., and David Moher. 2017. "Promote Scientific Integrity via Journal Peer Review." *Science* 357 (6348):256-7.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in peer review." *Journal of the American Society for Information Science and Technology* 64 (1):2-17.

Livneh, Ben, Eric A. Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M. Andreadis, Edwin P. Maurer, and Dennis P. Lettenmaier. 2013. "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions." *Journal of Climate* 26 (23):9384-9392.

Lozano, George A., Vincent Larivière, and Yves Gingras. 2012. "The weakening relationship between the Impact Factor and papers' citations in the digital age." *Journal of the American Society for Information Science and Technology* 63 (11):2140-45.

Macrae, C. Neil, Galen V. Bodenhausen, and Alan B. Milne. 1995. "The Dissection of Selection in Person Perception: Inhibitory Processes in Social Stereotyping." *Journal of Personality and Social Psychology* 69 (3):397-407.

Merton, Robert K. 1968. "The matthew effect in science." *Science* 1968:56-63.

Merton, Robert K. 1973. "The normative structure of science." In *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Norman W. Storer, 267-78. Chicago, IL: University of Chicago Press.

Mindner, Justin R., Philip W. Mote, and Jessica D. Lundquist. 2010. "Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains." *Journal of Geophysical Research: Atmospheres* 115. doi: https://doi.org/10.1029/2009JD013493.

National Academies of Sciences, Engineering, and Medicine,. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, D.C.: The National Academies Press.

National Science Foundation. 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. In *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*.

Nature Publishing Group. 2015. "Author Insights 2015 Survey."

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Mahlhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. "Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility." *Science* 348 (6242):1422-5. doi: 10.1126/science.aab2374.

Okasha, Samir. 2016. "On the interpretation of decision theory." *Economics & Philosophy* 32 (3):409-33.

Ormans, Laurent. 2016. "50 Journals used in FT research." accessed September 1. https://www.ft.com/content/3405a512-5cbb-11e1-8f1f-00144feabdc0.

Ridgeway, Cecilia L., and Shelley J. Correll. 2006. "Consensus and the creation and status beliefs." *Social Forces* 85 (1):431-53.

Rubin, Hannah, and Cailin O'Connor. 2018. "Discrimination and Collaboration in Science." *Philosophy of Science* 85:380-402.

San Francisco Declaration on Research Assessment. 2013. "The San Francisco Declaration on Research Assessment (DORA)." accessed September 1. https://sfdora.org/read/.

Sauder, Michael, and Wendy Nelson Espeland. 2006. "Strength in numbers? The advantages of multiple rankings." *Indiana Law Journal* 81 (1):205-27.

Schimmack, Ulrich. 2015. "Replicability Ranking of 26 Psychology Journals." January 18. https://replicationindex.wordpress.com/2015/08/13/replicability-ranking-of-26-psychology-journals/.

Science. "Mission and Scope." accessed September 1. http://sciencemag.org/about/mission-and-scope.

Strevens, Michael. 2003. "The role of the priority rule in science." *Journal of Philosophy* 100 (2):55-79.

West, Jevin D., Theodore C. Bergstrom, and Carl T. Bergstrom. 2010. "The Eigenfactor Metrics™: A network approach to assessing scholarly journals." *College & Research Libraries* 71 (3):236-44.

Willer, Robb, Ko Kuwabara, and Michael W. Macy. 2009. "The False Enforcement of Unpopular Norms." *American Journal of Sociology* 115 (2):451-90.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. 2015. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.

Wilsdon, James, Judit Bar-Ilan, Robert Frodeman, Elisabeth Lex, Isabella Peters, and Paul Wouters. 2017. Next-generation metrics: Responsible metrics and evaluation for open science. Report of the European Commission Expert Group on Altmetrics. European Commission.

Zollman, Kevin J. S. 2018. "The Credit Economy and the Economic Rationality of Science." *The Journal of Philosophy* 115:5-33.

Zuckerman, Harriet, and Robert K. Merton. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System." *Minerva* 9 (1):66-100.

Pragmatism and the content of quantum mechanics

Peter J. Lewis

Draft – please don't quote

Abstract

Pragmatism about quantum mechanics provides an attractive approach to the question of what quantum mechanics says. However, the conclusions reached by pragmatists concerning the content of quantum mechanics cannot be squared with the way that physicists use quantum mechanics to describe physical systems. In particular, attention to actual use results in ascribing content to claims about physical systems over a much wider range of contexts than countenanced by recent pragmatists. The resulting account of the content of quantum mechanics is much closer to quantum logic, and threatens the pragmatist conclusion that quantum mechanics requires no supplementation.

1. Introduction

Quantum mechanics is, notoriously, a theory in need of interpretation. But there is very little agreement on what kind of interpretation it needs. That is, there is very little agreement concerning what the foundational problems of quantum mechanics *are*, and without such agreement, there is little hope for a consensus concerning what an acceptable solution to the problems might look like.

Here is a way to divide up the territory. We can distinguish between *descriptive* and *normative* questions concerning quantum mechanics. Descriptive questions concern what quantum mechanics *says*—the *content* of the theory, as expressed in textbooks and used in labs. Normative questions concern what quantum mechanics *should* say—and in particular, whether it should say something different from what it actually does say.

All parties to the debates over the foundations of quantum mechanics would agree, I think, that there is a legitimate descriptive question concerning the content of quantum mechanics. Even those philosophers and physicists who think that quantum mechanics wears its interpretation on its sleeve at least feel the need to correct the mistaken impressions of *other* philosophers and physicists concerning what quantum mechanics says. The normative question presupposes an answer to the descriptive one: some think quantum mechanics is just fine the way it is, others contend that it needs to be replaced or supplemented with something radically different, and in large part this difference in attitude depends on prior differences concerning the answer to the descriptive question.

As an illustration, consider a fairly standard narrative concerning the descriptive and normative questions. Descriptively speaking, quantum mechanics depends on a distinction between measurements and non-measurements: measurements follow one dynamical law, the collapse dynamics, and non-measurements follow a different dynamical law, the Schrödinger dynamics. Since these two dynamical processes are incompatible, a precise formulation of quantum mechanics requires a precise dividing line between measurements and non-measurements. Quantum mechanics nowhere provides such a thing—and indeed, it seems highly unlikely that a term like “measurement” could be given a physically precise definition. So

descriptively speaking, quantum mechanics is inadequate as a physical theory. On the basis of this measurement problem, Bell (2004, 213–231) recommends replacing quantum mechanics with either a pilot-wave theory or a spontaneous collapse theory. For similar reasons, Wallace (2012, 35) recommends replacing quantum mechanics with a many-worlds theory.[1]

But not everybody concurs. There are alternative narratives according to which quantum mechanics, descriptively speaking, is just fine as it is, and hence there is no normative pressure to supplement or replace it. One prominent version proceeds from the quantum logic of von Neumann (1936) and Putnam (1975) through to the quantum information theory of Bub (2016). According to this approach, quantum mechanics describes a non-classical event space—in terms of truth values, a non-Boolean algebra, and in terms of probability ascriptions, a non-simplex distribution. No-go theorems (arguably) show that it is impossible to construct a set of events obeying classical Boolean logic or classical Kolmogorov probability that reproduces the empirical predictions of quantum mechanics. The implication is that in quantum mechanics we have discovered something important about the fundamental event structure of the world. Seeking to replace or supplement quantum mechanics with a theory obeying classical logic and classical probability theory amounts to a quixotic attempt to impose a structure on the world that it manifestly does not have (Bub 2016, 222). The measurement problem, on this account, results from a mistaken demand for a dynamical explanation of the individual events in the quantum structure, when no such explanation is available (Bub 2016, 223)

[1] Wallace takes the many-worlds theory to be a precise statement of the content of quantum mechanics, rather than a replacement for it. I take up the question of whether the many-worlds structure is present in quantum mechanics as it stands in section 2.

This fundamental difference of opinion—between those who take the measurement problem seriously and those who regard it as a pseudo-problem—continues to divide the foundations of physics community today. Hence the descriptive question—the question of what quantum mechanics actually *says*—remains a pressing one. In this paper, I argue for a particular way of approaching the descriptive question. The methodology is the pragmatist one of Healey (2012; 2017) and Friederich (2015), but the answer to the descriptive question that results from following this methodology, I argue, differs in an important way from the answers that Healey and Friederich give. I conclude by assessing the consequences of this answer to the descriptive question for the normative question.

2. The descriptive question

So how should we approach the descriptive question? Consider a straightforward realist approach to the content of scientific theories. A theory, at least in physics, is typically expressed using a particular mathematical structure. The *state* of a physical system is generally identified with a mathematical entity that resides in a particular abstract space, and the *dynamics* of the theory tell us how that state evolves over time. So, for example, in many applications of classical mechanics, the state of a physical system can be represented by a set of vectors in a three-dimensional Euclidean space, and the dynamical laws of Newtonian mechanics tell us how the set of vectors evolves over time. The interpretation of the mathematics is fairly straightforward: the vectors represent the positions and momenta of point-like particles, and classical mechanics tells us how the properties of the particles change.

Such an approach can equally be applied to quantum mechanics (Albert 1996). According to quantum mechanics, the state of a physical system is identified with a complex-valued function defined on a configuration space—a space with three dimensions for each particle in the system. A dynamical law, the Schrödinger equation, tells us how this function, the wave-function, changes over time. Then by analogy with classical mechanics, the wave-function must be a representation of the physical properties of the quantum system as they change over time.

The continuity with classical mechanics in the above account is attractive, but there are surprising consequences. For an *N*-particle system, the wave-function is defined over a 3*N*-dimensional configuration space, and it cannot be represented without loss in a three-dimensional space. This has led some to conclude that a straightforward realist reading of quantum mechanics shows that the three-dimensionality of our physical world is illusory (Albert 1996). Furthermore, if we model a measurement using quantum mechanics, the wave-function ends up with components corresponding to each possible outcome of the measurement—not just one outcome, as is the case classically. This leads Everettians like Wallace (2012) to conclude that a straightforward realist reading of quantum mechanics shows that every possible outcome of a measurement actually occurs.

These conclusions might be right, but do they simply follow from close attention to the structure of quantum mechanics? There are reasons to be suspicious. As Healey (2017, 116) notes, conclusions of this kind depend on the assumption that the wave-function plays the same descriptive role in quantum mechanics as the position-momentum vectors play in classical mechanics. If this assumption is itself up for grabs in the interpretation of quantum

mechanics, then neither of these conclusions is warranted. But how do we adjudicate the question of whether the wave-function describes physical systems or whether it has some other, non-descriptive role? Is there a metaphysically neutral methodology that could be used to answer this question? Healey (2012; 2017) and Friederich (2015) think that there is.

3. Pragmatism

Consider an analogy. "Stealing is bad" has the same grammatical structure as "Cherries are red". But it is far from clear that both sentences should be taken as descriptive. In particular, badness, taken as a property of actions, seems like a queer kind of property, imperceptible and disconnected from the other properties of the action. Expressivists seek to dissolve the problem of the nature of badness by claiming that a sentence like "Stealing is bad" should be taken as expressive rather than descriptive—as expressing our attitude towards stealing. Pragmatists further coopt expressivism as a variety of pragmatism (Price 2011, 9). Pragmatists stress the variety of uses of language, noting that sentences with superficially similar form can be used in radically different ways. "Cherries are red" is used to describe a class of objects, whereas "Stealing is bad" is used to express our attitude towards a class of actions.

Pragmatism, then, enjoins us to pay close attention to how a sentence is *used* in order to find out what it means. Healey (2012; 2017) and Friederich (2015) each suggest that the pragmatist approach provides us with a metaphysically neutral methodology for probing the content of quantum mechanics. That is, we can look at how various quantum mechanical claims are used by physicists in order to determine what those claims mean. This strikes me as a welcome suggestion. In the rest of this section I present the conclusions of their pragmatist

inquiries; in the next, I consider whether the language use of physicists actually supports those conclusions.

Healey (2012) distinguishes between *quantum claims* and *non-quantum magnitude claims*. The former explicitly mention quantum states, quantum probabilities, or other novel elements of the theory of quantum mechanics. The latter are claims about the magnitude of a physical quantity that do *not* involve quantum states, quantum probabilities etc. In keeping with the pragmatist methodology, Healey bases this distinction on the way the two kinds of claims are used. Non-quantum magnitude claims are used in a straightforwardly descriptive way. But quantum claims are used in a different way: they are used, not to *describe* a system, but to *prescribe* a user's degrees of belief in various non-quantum magnitude claims.

As an example, Healey appeals to the Interference experiments of Juffmann et al. (2009), in which C_{60} molecules are passed through an array of slits and then deposited on a silicon surface. To derive quantum mechanical predictions for this experimental arrangement, quantum states are ascribed to C_{60} molecules. That is, quantum claims of the form "The molecule has state $|\psi\rangle$" are used, via the Born rule, to ascribe probabilities to claims concerning the various possible locations of the molecules on the silicon surface. These latter claims—of the form "The molecule is located in region R"—are non-quantum magnitude claims. The job of the non-quantum magnitude claims is to describe the physical system, but the job of the quantum claims is to prescribe degrees of belief in the non-quantum magnitude claims for an appropriately situated observer. In this respect Healey's approach is like the expressivist's in ethics: claims that have superficially similar grammatical forms have very different functions.

Another important strand in the pragmatist approach concerns the role of decoherence. After the C_{60} molecule hits the silicon surface, complicated interactions with the surface mean that the state of the molecule-environment system becomes approximately diagonal when written as a density matrix in the position basis. This in turn insures that the probabilities ascribed by the Born rule to various claims about the molecule's position closely obey the probability axioms. But before the molecule encounters the silicon surface, its state is a coherent superposition—a state that is not even approximately diagonal, and for which the Born rule does not ascribe probabilities to location claims that closely obey the probability axioms. For such a state, the Born rule does not prescribe appropriate degrees of belief in the non-quantum location claims, and so assertion of such claims prior to decoherence is not *licensed* by quantum mechanics. Decoherence, then provides a demarcation between situations in which it is appropriate to have a well-defined degree of belief in a non-quantum magnitude claim, and situations in which it is not.

The central finding of the Healey-Friederich pragmatist approach is that attention to the use of quantum mechanical language shows that claims about the quantum state of a system are not used to describe that system. Hence, we should not think of the wave-function as a representation of the physical properties of the quantum system as they change over time. This perspective has the advantage that the measurement problem does not arise: if the wave-function doesn't represent the system, then we don't have to worry that the dynamical laws for wave-function evolution are different for measurements and non-measurements. In fact, if the quantum state is prescriptive, then the difference between measurements and non-

measurements arises quite naturally: the results of measurements have a direct and obvious influence on what you should believe.

Hence the pragmatist approach provides a clear answer to the descriptive question: quantum mechanics, in itself, says *nothing* about the world. As Healey (2017, 12) puts it, "quantum theory has no physical ontology". Rather, quantum mechanics tells us what to believe about non-quantum ontology—about particles, or in the case of quantum field theory, about fields. Furthermore, this answer to the descriptive question suggests an answer to the normative question: since the measurement problem doesn't arise, there is no motivation for supplementing or replacing quantum mechanics with something else.

4. Actual use, counterfactual content

Thus far, I have said little about the evidence that backs up Healey's claims about how quantum claims and non-quantum magnitude claims are used. Indeed, direct evidence from the language use of physicists is likely to be unenlightening: that a claim is asserted in a given context provides no direct evidence concerning whether its content is descriptive or prescriptive.

To fill this gap, Healey appeals to an inferentialist account of the link between use and meaning derived from the work of Robert Brandom (2000): the meaning of a claim is identified with the set of material inferences it licenses. So by looking at the way a claim is used in licensing inferences, we can gain evidence about what it means. And here the distinction between prescriptive quantum claims and descriptive non-quantum magnitude claims seems to be well motivated. In the practice of physics, a claim about the quantum state of a system is

used to infer Born probabilities, and nothing more. If Born probabilities are taken to be rational degrees of belief, then the prescriptive content of a quantum claim exhausts its meaning.

A non-quantum magnitude claim, on the other hand, can license a wide variety of inferences. From the claim that a C_{60} molecule is located in a particular region of the silicon surface, we can infer that an electron microscope will produce an image of the molecule if directed at that region (Juffmann et al. 2009, 2). We can infer that if the silicon surface is left untouched for two weeks, the C_{60} molecule will remain in the same place (Juffmann et al. 2009, 2). Under suitable conditions, we can infer that the C_{60} molecule will emits photons; under different conditions, that it will act as a nucleation core for molecular growth (Juffmann et al. 2009, 3). In other words, the inferences licensed by the non-quantum magnitude claim support the interpretation that the meaning of the claim is descriptive rather than merely prescriptive.[2]

So there is a good case to be made, I think, that actual use supports the distinction between prescriptive quantum claims and descriptive non-quantum magnitude claims. But there is a further strand to the Healey-Friederich interpretation, namely that non-quantum magnitude claims are only licensed after decoherence. This claim, I think, does not stand up so well to scrutiny.

Consider C_{60} interference again. After the molecule has adhered to the silicon surface, the state of the molecule is decoherent, and the claim that the molecule has a particular

[2] There is a sense in which the meaning of *any* claim is prescriptive according to the inferentialist program: the claim about the location of the molecule licenses an inference to a certain *degree of belief* that the electron microscope will produce an image of it. But still, there is a reasonable distinction here: the quantum claim licenses inferences only via the Born rule, whereas the non-quantum magnitude claim licenses inferences via a huge variety of schema typical of small physical objects. The latter is just what it is for a claim to be descriptive.

location is licensed—that is, it is appropriate to associate a particular degree of belief with the claim, and if that degree of belief is high enough, it is appropriate to assert the claim. But before the molecule has adhered to the silicon surface, the state of the molecule is coherent, and no claim about the location of the molecule is licensed—it is not appropriate to associate a degree of belief with such a claim, or to assert it. Similar considerations apply to properties other than location.

This seems to fly in the face of actual use. For example, in the description of the C_{60} interference experiment, Juffmann et al. (2009, 2) assert that "all transmitted particles arrive with the same speed," and "about 110cm behind the source, the molecules encounter the first diffraction grating," apparently ascribing both speed and location to C_{60} molecules prior to decoherence. This doesn't seem to be an isolated incident: physicists routinely talk of preparing, selecting, spraying, shooting and trapping particles, ions and molecules, and this talk typically involves making claims about these objects prior to any eventual decoherence.

It is possible, of course, that this is just "loose talk", or an indirect way of making claims about the quantum state of the systems concerned. But given the frequency of such claims, and given the reliance of the pragmatist methodology on *use*, this seems like a shaky game to play. It would be better, all things considered, if such claims could be accommodated within the pragmatist interpretation, rather than explained away as anomalies.

But there are obvious barriers to licensing non-quantum magnitude claims prior to decoherence. As Friederich (2015, 79) notes, the Born rule is only "reliable" when applied to decoherent states, in the sense that only for such states are the numbers it produces guaranteed to closely obey the probability axioms. Given some reasonable assumptions about

rationality, it is plausible that numbers that do not closely obey the probability axioms could not be rational degrees of belief. Furthermore, Healey argues that asserting a non-quantum magnitude claim prior to decoherence is likely to be misleading. For example, suppose one asserts (with Juffmann et al.) that "about 110cm behind the source, the molecules encounter the first diffraction grating." One might infer from this that each molecule passes through exactly one slit in the grating, and hence that the presence of the other slits is irrelevant, and hence that there is no possibility of interference (Healey 2012, 745).

So the pragmatist approach seems to face a dilemma: either it fails to accommodate the actual language use of physicists, or it licenses misleading assertions and irrational degrees of belief. Isn't there another way? I think there is. Consider a mundane claim like "There is beer in the fridge." In typical contexts, an assertion of this claim licenses the inference that if you were to go to the fridge and open the door, you could take a beer and drink it. Of course, you might not actually do this; maybe you don't want a beer. That is, the inference here is a counterfactual one. A good deal of the inferential content of our assertions has this counterfactual character.

Now return to the quantum context. Consider again the claim that "about 110cm behind the source, the molecules encounter the first diffraction grating." What content could that claim have? If we broaden the notion of inferential content to include counterfactual inferences, then the content seems fairly clear: if we were to replace the first diffraction grating with a detector taking up the same region of space, then the Born rule would ascribe a degree of belief close to 1 to detecting the molecules.

How does the inclusion of counterfactual content avoid the barriers to licensing non-quantum magnitude claims prior to decoherence? Note that the counterfactual content of the claim about the molecules involves a counterfactual intervention on the system—a counterfactual measurement. The counterfactual measurement induces counterfactual decoherence. The Born probabilities are conditional on this intervention and the associated decoherence, so the Born probabilities for various position claims concerning the molecules are not, after all, unreliable, in the sense of violating the probability axioms.

Neither should there be any danger of being misled by an assertion that the C_{60} molecules encounter the grating, because the counterfactual conditions implicit in the content of that assertion are distinct from the conditions that actually obtain in the apparatus. That you *could* detect the molecules at the diffraction grating, given a different experimental arrangement, doesn't license the inference that there *is* no interference, given the actual experimental arrangement. Admittedly, though, this amounts to a weakening of the content of position claims from the classical case, as spelled out in the next section.

5. A happy convergence?

I have argued that non-quantum magnitude claims have assertible content in a far wider range of contexts than countenanced by Healey or Friederich. If there is some counterfactual intervention on a system that would produce decoherence in the basis defined by a given observable, then claims about the values of that observable have content. And since counterfactual interventions only have to be realizable in principle, this means that claims about the value of an observable for a system *generally* have content, whether or not the

system *actually* decoheres in the basis defined by that observable. This has the welcome consequence that the frequent assertions made by physicists about the properties of systems prior to decoherence are contentful.

A potential cost of such permissiveness about content is that the structure of this content is, in general, non-Boolean. Consider again a C_{60} molecule that is approaching the first diffraction grating, and consider an assertion of "The molecule passes through the leftmost slit". This assertion has content, on the proposed view, because in principle there is an intervention on the system that would produce decoherence in a basis defined by an observable that distinguishes which slit the molecule passes through. Still, assertion of the claim would not be appropriate, simply because there are many slits in the grating, so the Born rule ascribes it a low probability. The same goes for every other slit in the grating. Nevertheless, the assertion that "The molecule passes through the leftmost slit, or the second to the left, or…" is assertible, since the Born rule ascribes it a probability close to 1. The disjunction is assertible, but none of the disjuncts is assertible. Since assertibility is a surrogate for truth in the pragmatist context, this is equivalent to saying that the disjunction is true, but none of the disjuncts is true.

One might take this to be unacceptable on the pragmatist view—especially if you endorse an inferentialist pragmatism, as Healey does. From a disjunctive claim you can straightforwardly infer that at least one of the disjuncts is true. If the content of a claim is identified with the inferences that it licenses, then part of the meaning of the disjunctive claim about the C_{60} molecule is that some assertion of the form "The molecule went through slit *x*" is true. Hence my proposal about content threatens to violate the inferentialist account of

meaning. The pragmatist interpretation of Healey and Friederich avoids this problem by insisting that claims about systems have meaning only after suitable decoherence.

Of course, pragmatism is not necessarily tied to an inferentialist account of meaning. But even given inferentialism, there is arguably no real problem here. Physicists are *selective* in the inferences they draw: from the disjunctive claim, they don't infer that the C_{60} molecule goes through some particular slit, so they don't infer a lack of interference. But they do infer that the molecule will arrive at the silicon surface, that it might radiate a photon in flight, and so forth. That is, the inferences drawn by physicists from their claims about pre-decoherent systems suggest that the non-Boolean structure of those claims is already *built in* to the meanings associated with those claims and revealed in inference.

This suggests that close attention to the way non-quantum magnitude claims are actually used leads to a happy convergence between pragmatism and the quantum logical approach. Physicists assert claims about particles even when the state does not decohere, and such claims seem to be meaningful. But physicists are not inclined on that basis to draw all the inferences that a full Boolean structure to their claims would license. Quantum mechanics apparently weakens the meaning of many claims about pre-decoherent physical systems, but without rendering those claims meaningless.

6. The normative question

As a methodology for addressing the *descriptive* question of the content of quantum mechanics, the pragmatist approach seems entirely appropriate: look to the *use* of physicists to determine what the various claims involved in the theory mean. At the hands of Healey and

Friederich, this approach yields the important insight that while non-quantum magnitude claims are used to describe physical system, quantum claims are used to prescribe appropriate degrees of belief in non-quantum magnitude claims. But Healey and Friederich go further, in limiting the assertibility of non-quantum magnitude claims to contexts in which the quantum state is decoherent in the relevant basis. This, I have argued, cannot be squared with the actual use of such claims. I propose instead that non-quantum magnitude claims *generally* have well-defined content, understood in terms of a counterfactual intervention on the system. This change to the pragmatist approach means that it ends up looking a lot like the quantum logical approach that preceded it. Indeed, the pragmatist approach might be regarded as a *justification* for quantum logical claims concerning the content of quantum mechanics.

But where does all this leave the *normative* question concerning whether quantum mechanics is fine as it is, or whether it should be supplemented or replaced? Healey and Friederich argue that quantum mechanics is fine as it is: if quantum claims do not describe physical systems, then there can be no conflict between the way that quantum mechanics describes systems during measurements and the way it describes them during non-measurements. If there is no measurement problem, then there is no motivation to replace such a successful theory. If, as Healey (2017, 12) maintains, quantum theory "states no facts about physical objects or events," then there can be no requirement that we come up with an *explanation* of quantum facts and events.

However, I have suggested that quantum theory has more content than the pragmatists countenance. In one sense, I agree that quantum theory states no facts: a quantum claim, such as the attribution of a quantum state to a system, is not a description. But in another sense,

there are distinctive quantum facts, or at least facts with a distinctive quantum structure: non-quantum magnitude claims about pre-decoherent systems exhibit the non-Boolean structure characteristic of quantum mechanics. This is the sense in which quantum logic gets things right.

Notably, though, the proponents of quantum logic *also* often take the view that quantum logic dissolves the measurement problem (e.g. Putnam 1975, 186). But this dissolution is widely regarded to be a failure (e.g. Bacciagaluppi 2009, 65) Once one has admitted that the structure of true (i.e. assertible) claims for a quantum system is non-Boolean, the question of *how* the world manages to instantiate this structure becomes legitimate and pressing. A denial that any explanation is required looks suspiciously like instrumentalism. And since any answer to this question goes beyond quantum mechanics as it stands, the call for explanation involves a demand to supplement quantum mechanics, or to replace it with something more fundamental.

Of course, given the no-go theorems, the path to an explanation of the structure of quantum facts is by no means clear. But neither do the no-go theorems show that an explanation is *impossible* (Friederich 2015, 161).[3] If the foregoing is correct, then pragmatism is an excellent way to *expose* the foundational problems of quantum mechanics, but it is not a means to *dissolve* them.

References

[3] Interestingly, Friederich (2015, 161) suggests supplementing quantum mechanics with sharp values for all observables, even though this seems at odds with his therapeutic aim of dissolving the foundational problems of quantum mechanics rather than solving them (2015, 6).

Albert, David Z. (1996), "Elementary quantum metaphysics," in J. T. Cushing, A. Fine and S. Goldstein (eds.), *Bohmian Mechanics and Quantum Theory: An Appraisal*. Dordrecht: Springer, 277-284.

Bacciagaluppi, Guido (2009), "Is logic empirical?" in K. Engesser, D. M. Gabbay and D. Lehmann (eds.), *Handbook of Quantum Logic and Quantum Structures.* Amsterdam: North-Holland, 49-78.

Bell, J. S. (2004), *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press.

Brandom, R. (2000), *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Bub, Jeffrey (2016), *Bananaworld: Quantum Mechanics for Primates*. Oxford: Oxford University Press.

Friederich, Simon (2015), *Interpreting Quantum Theory: A Therapeutic Approach*. Basingstoke: Palgrave Macmillan.

Healey, Richard (2012), "Quantum theory: a pragmatist approach," *British Journal for the Philosophy of Science* 63: 729-771.

Healey, Richard (2017), *The Quantum Revolution in Philosophy*. Oxford: Oxford University Press.

Juffmann, T., Truppe, S., Geyer, P., Major, A. G., Deachapunya, S., Ulbricht, H., and Arndt, M. (2009), "Wave and particle in molecular interference lithography," *Physical Review Letters* 103: 263601.

Price, Huw (2011), *Naturalism Without Mirrors*. Oxford: Oxford University Press.

Putnam, Hilary (1975), "The logic of quantum mechanics," in *Mathematics, Matter and Method: Philosophical Papers Volume 1*. Cambridge: Cambridge University Press.

von Neumann, John (1932), *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer-Verlag.

Wallace, David (2012), *The Emergent Multiverse*. Oxford: Oxford University Press.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -436-

Tool Migration: A Framework for Analyzing Cross-disciplinary Use of Mathematical Constructs

Chia-Hua Lin
University of South Carolina / KLI

Abstract. Mathematical formalisms that are constructed for inquiry in one disciplinary context are sometimes applied to another, a phenomenon that I call 'tool migration.' Philosophers of science have addressed the advantages of using migrated tools. In this paper, I argue that tool migration can be epistemically risky. I then develop an analytic framework for better understanding the risks that are implicit in tool migration. My approach shows that viewing mathematical constructs as tools while also acknowledging their representational features allows for a balanced understanding of knowledge production that are aided by the research tools migrated across disciplinary boundaries.

Keywords: Cross-disciplinarity, tool migration, epistemic risks

1. Introduction

Mathematical formalisms that are constructed for scientific inquiry in one disciplinary (or sub-disciplinary) context are applied to another. Philosophers of science have started paying attention to this cross-disciplinary aspect of scientific practice. For instance, the discussion of 'model transfer' concerns a relatively small set of mathematical models that are applied in multiple disciplinary contexts. Humphreys (2004) proposes that models that are transferred to study phenomena of a different domain owe their versatility to the computational tractability they afford. In contrast, Knuuttila and Loettger (2014, 2016) suggest that in addition to tractability, versatile models also offer conceptual frameworks for theorization, which they label 'model templates.' However, these analyses do not deal with the risks inherent in this aspect of scientific practice. Consider the use and development of game theory in evolutionary biology as an example. In importing game theory, which was originally conceived to describe strategic interaction between rational agents typically studied by social scientists, evolutionary biologists may need to modify the theory in order to generate knowledge about presumably non-rational agents, at least in many cases. One can then assume that any changes to the theory--between its established applications in social sciences and its novel uses in evolutionary biology--require special attention so as to avoid misinterpreting an analysis.

Despite the advantages, there might be risks associated with using mathematical constructs across disciplines. In this paper, I ask: might there be patterns of transfer that may undermine the effectiveness of the imported mathematical formulation? What would these

patterns, if any, look like? This paper is an attempt to explore the conditions in which importing mathematical constructs may be epistemically risky. To begin, I develop a framework to systematically characterize the landscape of mathematical importations. The goal of such a framework is two-fold. Proximately, the framework captures characteristics of migration that the current terminology, such as 'model transfer' or 'importing/exporting,' fails to discern. Ultimately, with this additional discernibility, I suggest that one may start to explore and identify patterns of importation that may be subject to epistemic risks, such as misinterpretation of an outcome produced by using a imported mathematical construct.

In Section 2, I argue that one can view mathematical constructs in science in terms of 'research tools' and that transporting such tools across disciplines, which I call 'tool migration,' can in some cases be a disservice to science. Next, I classify tool migration based on two kinds of contextual details that bear significance to the effectiveness of the migrated research tool in a foreign context. In Section 3, I apply this approach to the use and development of game theory in evolutionary biology. Finally, in Section 4, I discuss in what ways this tool migration framework, which is essentially a typology of four types of tool migration, may help to characterize epistemically risky patterns of tool migration.

2. Theoretical Background

Although the notion of epistemic risks associated with migration of mathematical constructs has not been explicitly addressed, the idea of viewing mathematical constructs as research tools follows from the discussion on the ontology of scientific models. Ever since the shift of attention to scientific practice (e.g., Hacking 1983), there has been a growing literature in which models in science are viewed as entities *detachable* from theory and data (e.g., Morrison 1999; Morgan and Morrison 1999). One recent predecessor to my tool migration account is a pragmatic approach to scientific models put forth by Boon and Knuuttila (2008). In their paper, which uses examples from engineering, they argue that scientific models are better understood as 'epistemic tools' instead of as representations of some target systems in the world. Boon and Knuuttila's argument draws heavily on the epistemological roles of scientific models in relation to the scientists who use them. According to them, scientific models allow their users "to understand, predict, or optimize the behavior of devices or the properties of diverse materials" (2008, 687). Thus, for an ontological account of scientific models to be productive and realistic, as they argue, it should be sensitive to the relation between the models and the modelers, i.e., the tools and their users. An adequate evaluation of Boon and Knuuttila's argument will take us far afield, but my work will show that both the representational and the pragmatic aspects are indispensable to a better understanding of the epistemic risks in tool migration.

2.1 Viewing mathematical constructs as research tools

In general terms, any mathematical construct that is to be *used or operated* in an algorithmic manner, and the outcome of whose operation is to be *interpreted* in order to answer a research question, is an example of what I am calling a research tool. Let me first unpack the operational aspect of a research tool.

Let's assume that the proper use of any mathematical constructs employed in scientific research is expected to produce consistent results. To achieve this consistency, then, a well-defined procedure needs to accompany such a construct so that anyone who follows the procedure expects, and is expected, to obtain the same outcome given the same input. For instance, when performing a game-theoretic analysis, one goes through a sequence of steps, such as: (i) identify the players and the acts available to them, (ii) identify the payouts in every set of acts, (iii) find the 'Nash equilibria,' which refers to a set of acts, one for each player, in which no player could improve his or her payoff by unilaterally changing act. A similar algorithmic procedure can be seen when applying, say, Newton's law of gravitation:

$$F_{grav} = G\frac{m_1 m_2}{r^2}. \qquad (1.1)$$

For example, the sequence of steps to obtain the magnitude of the gravitational force, F_{grav}, between any two objects includes: (i) identify the mass of each object, (ii) identify the distance between them, (iii) complete the equation in which 'm_1' and 'm_2' refer to the masses of the two objects, 'r' the distance in between, and 'G' the gravitational constant. In these two examples, when the first two steps produce consistent input, the third step is expected to generate the same output.

Moreover, concerning the interpretational aspect of a research tool, the output of a series of symbol assignments and manipulations can be understood *only through the lens of some interpretation*. The Nash-equilibrium of a game is a meaningful 'solution' in virtue of the usual understanding of the game-theoretic formulation of a problem. Similarly, the meaning of the value obtained through completing the equation in (1.1) is derived from the usual interpretation of the quantities appearing in the equation and the theoretical context in which those quantities are defined.

Finally, assume that something can be viewed as a tool if it serves as a means to an end. In this case, then, mathematical constructs like game theory or mathematical formulas can be seen as research tools. In the case of applying a mathematical construct, the goal of performing a sequence of prescribed steps goes beyond merely completing the calculation and obtaining a result. Instead, the output is to be interpreted so that one may solve a problem, answer a research question, or gain knowledge about a subject-matter. Thus, a mathematical construct that prescribes algorithmic symbol manipulation can be seen as a research tool, assisting its users to meet an end. Manipulating symbols is a means to the end that was specified during the mathematical formulation of the research problem.

2.2 Epistemic risks of tool migration

Another predecessor to my account is Morgan's discussion of the re-situating of knowledge (2014). According to her, knowledge production is necessarily 'situated,' and consequently, applying a piece of knowledge outside its initial context requires effort - different contextual situations require different 're-situating' strategies. The term 're-situation' thus captures what scientists do in practice to transport locally generated knowledge across contexts. As she argues, to make an instance of scientific knowledge accessible outside its production site, one needs to establish inferential links between the production site and the destination site. However, she suggests, whether a re-situation of knowledge contributes to scientific progress depends on whether the transport secures some sort of inferential safety.

Building from Morgan's notion of the re-situation of knowledge, I argue that cross-disciplinary use of research tools is epistemically risky. Given the locality of scientific knowledge production, applying scientific knowledge outside its production site may come with epistemic risks. For example, between the production site and a destination site, there may be incongruent disciplinary characteristics (e.g., implicit theoretical assumptions) that fail to be captured by the inferential strategy, such that knowledge from the former cannot be transferred to the latter. Similarly, we can assume that the construction of a research tool is also *situated* in nature. Namely, a research tool is conceived to be operated and to extend our knowledge concerning a subject-matter *given a particular disciplinary context*. It follows that cross-disciplinary use of research tools is as epistemically risky as re-situating knowledge. That is, the epistemic reliability (i.e., general ability or tendency to produce knowledge) of some research tool in one disciplinary context does not necessarily carry over to another.

The concept of 'tool migration' captures both the 'situated-ness' of a research tool that was established in its native discipline and the effort it takes to 're-situate' the tool in a foreign discipline. Naturally, in the process of uprooting a research tool, significant contextual details—ranging from implicit expertise to important background assumptions—may be stripped away. Likewise, during re-situation, new features may be introduced to the tool so as to treat a different subject matter in a new disciplinary context. Together, due to the possibility of losing or gaining significant contextual details, or both, a cross-disciplinary tool migration risks undermining the effectiveness of the tool. These risks include, for example, misinterpretation of the research result or failure to produce genuine knowledge. Thus, it follows that tool migration can in some cases be a disservice to the production of knowledge.

Acknowledging these challenges, some have argued against the cross-disciplinary effort to integrate disciplinary knowledge (e.g., van der Steen 1993). Alternatively, one might try to overcome these challenges so long as the risks are better understood and managed. To understand the risks, I suggest that we first look at the patterns of tool migration. Among these patterns, we might find that some of them could be epistemically risky. Having established the

notions of research tools and risks involved with tool migration, I turn to the contextual details that are closely related to a tool's epistemic performance.

2.3 Contextual details of a research tool: the target profile and the usage profile

The construction of a research tool is necessarily situated within a context. In order to compare and contrast between the native (or established) context and the foreign context of a migrated tool, I single out two major types of details.

The first type concerns the assumptions about the entities that are studied by a subject-matter for which the tool is developed. For instance, game theory defines what it considers as a game, a player, or an act. For simplicity, I call *all* the assumptions that a tool makes about its target entities the tool's 'target profile.'

The second type considers *the ways* in which one interprets the output from applying a tool in his or her research. In a game-theoretic analysis, for example, by following an algorithmic procedure, one obtains a solution of a game in the form of a Nash equilibrium. Depending on the game that one was analyzing, the solution could be understood as an explanation of economic behavior, or a prediction about it, or it could be used to optimize an strategic interaction. For simplicity, I call *all* the ways in which a tool is intended to be used, e.g., describing, predicting, optimizing, or explaining its target phenomenon, the tool's 'usage profile.'

Together, as I demonstrate in Section 4, the 'target profile' and 'usage profile' allow one to detect patterns of changes in the contextual details between the established use and the novel use of a research tool. They are able to do this because these two profiles offer a coarse resolution; looking through the lens of the target profile and usage profile, one zooms out from particular cases of tool migration so as to detect patterns of cross-disciplinary transport. Further analyses of these patterns will then shed lights on their associated epistemic risks.

2.4 Four types of tool migration

With the two profiles of a research tool and the two contexts in which the tool is used, i.e., a novel use and an established use, one can distinguish four types of tool migration.

First, compared to its established use, when a novel use of a tool catalyzes changes in both target and usage profiles, the tool migration is transformative, and therefore I call it a ***tool-transformation***. Second, in contrast, when both target and usage profiles remain more or less intact after the migration, the tool's novel use is considerably similar to its previous applications. Thus, I call such a case ***tool-application***. Between these two extreme types, there are novel uses of a research tool that alter only one of the two profiles but not both. When a tool changes its target profile but not its usage profile, I call it a ***tool-transfer,*** and when a tool changes its usage profile but not the target profile, I call it a ***tool-adaptation***. See **Table 1** for a summary.

Table 1
A Typology of Tool Migration

Between established and novel uses of a research tool	**Usage profile remains**	**Usage profile deviates**
Target profile remains	'Tool-application'	'Tool-adaptation'
Target profile deviates	'Tool-transfer'	'Tool-transformation'

Among these four types of tool migration, tool-transfer is arguably the most familiar to the philosophers of science. Humphreys coins the term 'computational templates' to refer to a relatively small number of mathematical equations that are applied to investigate different domains of phenomena (2002, 2004). Bailer-Jones (2009) discusses such a scientific practice in terms of mathematical analogy. For one example, Newton's law of gravitation was intentionally sought after to model electrostatic force (see Bailer-Jones 2009 for a detailed account). The important parallel between the two formulas, shown in (1.2), is that both types of forces (gravitational and the electrostatic) are proportional to the inverse of the square of the distance, r, between two masses, m_1 and m_2, or two charges, q_1 and q_2. The constants that appear in both formulas scale the quantities to match empirical phenomena.

$$F_{grav} = G\frac{m_1 m_2}{r^2} \quad \text{and} \quad F_{el} = k\frac{q_1 q_2}{r^2} \tag{1.2}$$

In contrast, the other three types of tool migration, despite prominent examples, are less explored in regard to their general features. One prominent example of tool-transformation is the development of game theory to be used in evolutionary biology.

3. The Migration of Game Theory From Social Sciences to Biology

In this section, I show in what sense the novel use of game theory in evolutionary biology, which is now known as 'evolutionary game theory' ('EGT') can be considered as a tool-transformation. I should mention that my account of the migration of game theory in this paper is not meant to address all the limitations of both game theory and EGT in their respective disciplinary contexts. Instead, the purpose of this account is to show that one *can* detect patterns of migration that have epistemic implications by focusing on the target profile and usage profile of a research tool.

3.1 Game theory in social sciences

Game theory was initially formulated to mathematically model strategic interactions between intelligent, rational agents. In game theory, a game is defined as an interaction between two or

more players in which each player's payoff (e.g., profit) is affected by the decisions made by other players. Typically, such a game assumes both *perfect information* and *common knowledge. Perfect information* assumes that all players know the entire structure of the game (all moves and all payouts) as well as all previous moves made by all players in the game (if it is an iterated or multi-move game). *Common knowledge* is the assumption that all players know that all players have perfect information, and that all players know that all players know that all players have perfect information, and so on. That is, *common* knowledge concerns what players know about what other players know. Moreover, the players also recognize that all players are cognizant that all players are rational, i.e., there is common knowledge of the game and of the *unbounded rationality* of all players. As such, all players will act in the way that takes all other players' potential moves into account in order to maximize their odds of winning. In addition to these assumptions regarding the players of a game, the structure of a game, which refers to the combinations of each move and its payout, is usually summarized in a 'payoff matrix.' Typically, an analysis of a game aims to find out its 'solution,' a unique Nash equilibrium (or sometimes equilibria) of the game.

Game theory has been used in economics, as well as other social sciences, to describe, predict, optimize, or explain a variety of human interactions, such as the economic behaviors of firms, markets, and consumers (e.g., Brandenburger and Nalebuff 1995; Casson 1994) military decisions (Haywood 1954) or international politics (e.g., Snidal 1985).

3.2 Game theory in evolutionary biology

Game theory was later used in evolutionary biology, where a game is understood as phenotypes (or heritable traits) in contest. In 1973, John Maynard Smith and George Price borrowed the formalism of a payoff matrix from game theory to mathematically model the evolution of phenotype frequencies in a population of organisms (see Grüne-Yanoff 2011). Their modeling method assumed that phenotypes are in contest with other phenotypes in a population of organisms. For instance, in a Hawk-Dove game, the contest is embodied by organisms with the phenotype of being aggressive and other organisms that are peaceful. In such a context, the payoff of a move is interpreted as the reproductive success of the phenotype (i.e., the number of copies it will leave to the next generation). Moreover, while the terminology such as 'game,' 'payoffs' and the formalism of a payoff matrix can be seen in the novel use of game theory in biology, the solution to a game in evolutionary biology is decidedly different from the Nash-equilibrium. An evolutionary game theoretic analysis typically looks for an evolutionarily stable strategy (ESS), i.e., a distribution of phenotypes in a population that is stable.

3.3 Epistemic implications of tool transformation

It is clear that the target profile of game theory is no longer the same between its established use

in social sciences and its novel use in biology. First, none of the assumptions of *perfect information*, *common knowledge*, and *unbounded rationality* in what is now known classical game theory (CGT) remain in the novel use of game theory in biology. Second, the moves in EGT are heritable phenotypes exhibited by a group of organisms instead of acts available to players. Third, the payoffs in EGT are the reproductive success of the heritable traits. In this sense, the three assumptions concerning the players were stripped away from the tool - as a result of uprooting game theory from social sciences, and the *heritability* assumption about the moves as well as Darwinian fitness interpretation of the payoff were introduced to the tool - as a result of re-situating it to evolutionary biology.

Note that the change in the target profile forces a limitation to the usage profile of the migrated tool. For instance, nullifying the *unbounded rationality* assumption concerning the players, EGT can no longer be used to optimize a game, i.e., discovering the rationally optimal strategy, which is a common use of game theory in social sciences. For instance, in the prisoner's dilemma, the Nash-equilibrium is for both players to defect. This solution is often interpreted as a prescription for the game; the players are irrational not to defect. However, in a Hawk-Dove game, the ESS obviously has no such normative use. Because the 'moves' of being an aggressive type or a peaceful type are not 'chosen,' the idea of there being normatively better or worse choice of moves is therefore questionable. Moreover, the organisms are not assumed to be rational. Thus, while the players in the prisoner's dilemma could be said to be irrational for choosing to cooperate, this sense of normativity does not carry over to the evolutionary game theoretic analysis of the Hawk-Dove game. One would be mistaken to say that it is 'irrational' for the doves to be doves. Thus, the change in the target-profile of game theory, especially the stripping away of the *unbounded rationality* assumption, has resulted in how the migrated tool should or should not be used.[1]

Moreover, applying EGT to study social phenomena (e.g., Axelrod 1984) or cultural evolution (e.g., Skyrms 2010) requires a careful re-defining of the terms (such as fitness) so as to avoid misinterpretation. Using EGT in social sciences, which can be considered as a 'homecoming' of the migrated tool, is not uncommon. However, the notion of payoffs in EGT refers to, roughly, the overall biological reproductive success of a group of organisms that exhibit a phenotype. Obviously in a social context, reproductive success of the members of some group is not, very often, the feature of interest. A careful reinterpretation of payoffs is thus needed in every analysis to prevent misleading conclusions.

[1] Of course, a more interesting prescriptive use of the ESS of a Hawk-Dove game might be, for example, to manage ecosystems for optimal predator-prey balance. Nevertheless, it should be noted that a justification for this type of prescriptive use of EGT would require further analysis because it is apparently not be derived from CGT.

To generalize, this example suggests that at least in some cases, a change in the target profile requires a corresponding change in the usage profile, or failure of producing genuine knowledge may follow. So far, I have shown that a solution of an ESS analysis may not be interpreted as an optimization to a Hawk-Dove game. Applying EGT to study social phenomena also requires careful treatment to the notion of payoff. Now if, hypothetically, some researcher were to make either of these two mistakes, his or her novel use of the tool would have been classified as tool-transfer - the novel use changes only the target profile without also changing the usage profile. It suggests that in some cases, tool-transformation may not be as risky as tool-transfer. I will come back to the issue of tool-transfer after some remarks related to the migration of game theory.

4. Contributions of the Tool Migration Analysis

The tool migration typology and its focus on tracking both similarities and differences meets the needs to sharpen discussions concerning inter- or cross-disciplinary use of research tools. Current literature seems to lack a framework to capture important, relational characteristics of the research tools that appear in multiple disciplinary contexts. For instance, 'tool-transformation' captures significant differences in details between CGT and EGT without losing sight of the contextual relationship between the two. In contrast, other terms in the literature, such as 'imports' or 'transfers,' fall short of doing so.

'Imports' signals the importation of research tools from a foreign discipline. In contrast, 'transfers' refers to the use of a scientific model, which was established to study phenomena of one domain, to study phenomena of a different domain. Neither term captures the migration of game theory to biology. As Grüne-Yanoff argues,

> [B]iologists constructed the more sophisticated formal [evolutionary game theoretic] concepts themselves. One could speak of the import of formal concepts only with respect to very basic notions such as strategies or pay-off matrices, and it may be more appropriate to refer to formal inspirations rather than imports or transfers in these contexts. (2011, 392)

Moreover, I have suggested that a change in a tool's target profile without a corresponding change in the tool's usage profile *may* lead to misinterpretation and hence misuse of the tool. If this observation is generalizable, which is debatable, then it follows that cases of tool-transfer are epistemically riskier than cases of tool-transformation. On the other hand, if this observation applies only to some cases, it nevertheless reveals at least two epistemic implications concerning tool migration: 1) when the target profile changes, one must be careful not to draw conclusions that might be natural in the old context but may not make sense within the new context, given the new target, and 2) sometimes a change in target profile can, force a change in usage profile. Potentially failing to recognize when these changes occurred in a migration leads

to risky uses of the migrated tool.

Morgan (2011) has argued that while not all scientific knowledge travels far, those that travel with integrity (i.e., maintaining their content more or less intact during its travels) and travel fruitfully (i.e., finding new users or new functions) are considered to be traveling well. It is relatively easy to quantify the latter feature – one needs to look at just the number of a tool's novel applications. However, determining whether a tool has traveled with integrity is not straightforward. As a starting point, this proposed tool migration framework—especially its distinction between the target profile and the usage profile of a tool—provides a starting point that is crucial for assessing the integrity of a migrated research tool. With this framework, one may discover more patterns of tool migration that impact the epistemic integrity and, consequently, effectiveness of a migrated research tool in a foreign discipline.

5. Conclusion

I have argued that mathematical constructs used in science can be viewed as research tools and their cross-disciplinary novel use as tool migration. I have also argued that making novel use of established tools has its risks, but such an implication is not meant to deter cross-disciplinary sharing of tools. Indeed, certain important breakthroughs in the history of science are due to creative, unconventional, uses of research tools (e.g., the use of Fourier's mathematical treatment of heat to study electrostatics [Thomson 1842] or the use of Faraday's mechanical model of fluid motion to model the electromagnetic field [Maxwell 1861]). Versatile research tools are not rare in science. A framework of tool migration aims to offer not only a useful terminology to characterize the diverse landscape of their versatility but also a groundwork to investigate risky patterns of making novel use of established research tools. Finally, this tool migration approach shows that viewing these constructs as tools whilst acknowledging their representational features (i.e., as captured in their target profile) allows for a balanced understanding of knowledge production - especially those productions that are aided by research tools that have migrated across disciplinary boundaries.

References

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Bailer-Jones, Daniela M. 2009. *Scientific Models in Philosophy of Science*. University of Pittsburgh Press.

Brandenburger, Adam M., and Barry J. Nalebuff. 1995. *The Right Game: Use Game Theory to Shape Strategy*. Harvard Business Review.

Boon, Mieke, and Tarja Knuuttila. 2009. "Models as Epistemic Tools In Engineering Sciences: A Pragmatic Approach." In *Handbook of the Philosophy of Science*, edited by Anthonie Meijers, 687–720. Elsevier B.V.

Casson, Mark. 1994. *The Economics of Business Culture: Game Theory, Transaction Costs, and Economic Performance*. Oxford University Press.

Grüne-Yanoff, Till. 2011. "Models as Products of Interdisciplinary Exchange: Evidence from Evolutionary Game Theory." *Studies in History and Philosophy of Science Part A* 42 (2): 386–97.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Haywood Jr, O. G. 1954. "Military Decision and Game Theory." *Journal of the Operations Research Society of America* 2 (4), 365–85.

Humphreys, Paul. 2002. "Computational Models." *Philosophy of Science* 69 (September): 1–27.

___ 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Knuuttila, Tarja, and Andrea Loettgers. 2014. "Magnets, Spins, and Neurons: The Dissemination of Model Templates across Disciplines." *The Monist* 97 (3). The Oxford University Press: 280–300.

Knuuttila, Tarja, and Andrea Loettgers. 2016. "Model Templates within and between Disciplines: From Magnets to Gases–and Socio-Economic Systems." *European Journal for Philosophy of Science* 6 (3). Springer: 377–400.

Maynard Smith, John, and George Price. 1973. "The Logic of Animal Conflict." *Nature* 246: 15–18.

Maxwell, James Clerk. 1861. "Xxv. on Physical Lines of Force: Part I.–the Theory of Molecular Vortices Applied to Magnetic Phenomena." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21 (139): 161–75.

Morgan, Mary, and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Vol. 52. Cambridge University Press.

Morgan, Mary. 2010. "Travelling Facts." In *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, edited by Peter Howlett and Mary Morgan, 3–39. Cambridge University Press.

___ 2014. "Resituating Knowledge: Generic Strategies and Case Studies." *Philosophy of Science* 81 (5). University of Chicago Press: 1012–24.

Morrison, Margaret. 1999. "Models as Autonomous Agents." In *Models as Mediators: Perspectives on Natural and Social Science*, edited by Mary Morgan and Margaret Morrison, 38–65. Cambridge University Press.

Skyrms, Brian. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press.

Snidal, Duncan. 1985. "The Game Theory of International Politics." *World Politics* 38 (1). Cambridge University Press: 25–57.

Thomson, William. 1842. "On the Uniform Motion of Heat in Homogeneous Solid Bodies and Its Connection with the Mathematical Theory of Electricity." *Cambridge Mathematical Journal* 3 (1842): 71–84.

Van Der Steen, Wim J. 1993. "Towards Disciplinary Disintegration in Biology." *Biology and Philosophy* 8 (3): 259–75.

Seattle, WA; 1-4 November 2018 -447-

Representations are Rate-Distortion Sweet Spots

Manolo Martínez (mail@manolomartinez.net)

Abstract

Information is widely perceived as essential to the study of communication and representation; still, theorists working on these topics often take themselves not to be centrally concerned with "Shannon information", as it is often put, but with some other, sometimes called "semantic" or "nonnatural", kind of information. This perception is wrong. Shannon's theory of information is the only one we need.

I intend to make good on this last assertion by canvassing a fully (Shannon) informational answer to the metasemantic question of what makes something a representation, for a certain important family of cases. This answer and the accompanying theory, which represents a significant departure from the broadly Dretskean philosophical mainstream, will show how a number of threads in the literature on naturalistic metasemantics, aimed at describing the purportedly non-informational ingredients in representation, actually belong in the same coherent, purely information-theoretic picture.

1 Information, Shannonian and Dretskean

In what follows I will use a random variable, S, to encode the state the world is in, and another random variable, M, for signals. How should we characterize the information that values of M (i.e., individual signals) carry about values of S (i.e., individual world states)? The most basic quantity with which information theory records dependence among two random variables is the *mutual information* between them. This quantity being an expected value, Dretske (1981, p. 52f) claims, renders it unsuitable for an analysis of representational status, and it should be substituted by notions that record relations between individual states, S_i, and individual signals, M_j. The basic relation which substitutes mutual information in contemporary Dretskean accounts is that of *making a probabilistic difference* (Scarantino 2015): a signal M_j makes a probabilistic difference to the instantiation of a state S_i iff the following *basic inequality* holds:

$$P(S_i|M_j) \neq P(S_i)$$

Nearly all the accounts of information developed in the recent, and not so recent, philosophical literature on this topic are variations on, and attempts to quantify, this inequality. For illustration, in Skyrms (2010, p. 36) the "information in $[M_j]$ in favor of $[S_i]$" is defined as the *pointwise mutual information* (Also *pmi* henceforth) between

state and signal. There is a direct relation between pmis and the basic inequality: the former are nonzero iff the latter is true.

The running thread connecting most prominent contemporary accounts of information is that all there is to Shannon's information theory, at least for the purposes of investigating the nature of representation, is two quantities: the unconditional probability of states and the probability of states conditional on signals, perhaps rearranged as the logarithm of their ratio, or in some other way. Unsurprisingly, from this it is routinely concluded that there is much more to representation than information. This conclusion is premature: informational content in the Dretskean tradition is not by a long shot all there is to information theory. This should not be taken to imply that information is all there is to representation—for one thing, I believe with teleosemanticists (Millikan 1984; Papineau 1987) that teleofunctions have a role to play in a complete theory of representation—but it does mean that no Dretske-style "semanticized information" needs to be recognized, over and above the quantities studied in information theory proper. I will argue that it also means that some prominent proposals as to ways to bridge the information-representation gap are, in fact, unwittingly appealing to informational structure.

In the following section I review two such proposals. My aim is not to argue against them—they are built upon largely correct insights. I will instead aim at showing that a better informed understanding of information provides a way to incorporating these insights in a unified, purely information-theoretic picture.

2 Bridging Information and Representation

2.1 Many-to-One-to-Many Architectures

The first proposal is that it is not enough that representations carry information; on top of that, they must sit in the right place in a certain cognitive architecture. Sterelny (2003), for example, has argued that the emergence of representations is enabled by two prior evolutionary transitions: from "detection" to "robust tracking", on the one hand; from "narrow-banded" to "broad-banded" behavioral responses, on the other. Robust tracking is in essence a *many-to-one* relation between world state and signal: many sensory inputs give rise to one and the same representation. Other theorists have advocated similar architectural constraints on representational vehicles. Famously, Burge (2010) places a great deal of weight on *perceptual constancies* in his characterization of perceptual representation (Burge 2010, p. 413.) This is a variation on Sterelny's idea and, as such, a many-to-one architectural constraint on representational status.

As for broad-banded responses, in these systems a single representation will be flexibly dealt with, resulting in different courses of action, depending on the context where the representation is tokened. Response breadth is in essence a *one-to-many* relation between representational vehicle and output: one representation, many agential outputs.

2.2 Reference Magnetism

A second proposal has been to focus on the entities that should figure in the content of simple representations. The suggestion, typically, is that represented entities should be appropriately *natural*, or *real*. For example, Dan Ryder (2004, 2006) has argued that neurons become attuned to *sources of correlation*. These entities are closely related to Richard Boyd's *homeostatic property clusters* (also HPC henceforth, Boyd 1989): HPC theory identifies natural kinds with clusters of properties which tend to be instantiated together, and such that this frequent co-instantiation is not just a statistical fluke. What Ryder calls sources of correlation are the grounds for these HPC-related frequent co-instantiations—whatever it is that makes them *not* statistical flukes. Ryder claims that many of the representations the brain trades in target sources of correlation. Martínez (2013) and Artiga (forthcoming) have made more general cases that simple representations preferably target HPCs (Martínez), or properties that best explain the co-occurrence of other properties (Artiga).

A similar idea has been explored in an entirely independent line of enquiry starting with Lewis (1983): "among the countless things and classes there are ... [o]nly an elite minority are carved at the joints, so that their boundaries are established by objetive sameness and difference in nature. Only these elite things and classes are eligible to serve as referents" (Lewis 1984, p. 227). This is what Sider (2014, p. 33) calls *reference magnetism*.

As I show in section 4, these two ideas, although apparently disparate, are in fact closely related, and the explanatory payback they bring to representation-involving talk depends on their informational underpinnings.

3 Information Theory is a Source-Channel Theory

Philosophy has understood information theory as a mostly *definitional* effort: for all philosophers have typically cared, the theory begins and ends with a presentation of what it takes for one random variable (or the worldly feature it models) to carry information about another. But information theory goes well beyond that. It is, well, a *theory*, and as such it is chiefly composed of claims that are advanced in the hope that they be true about the world.

In a nutshell, the most celebrated results in information theory have to do with specifying how faithful the transmission of information from a source can be, when it happens over a (typically noisy, typically narrow) channel. These results have played absolutely no role in informational accounts of representation.[1] Take, for starters, the idealized depiction of an information-processing pipeline in fig. 1 (*cf.* Cover & Thomas 2006, fig. 7.1)

[1] Two recent philosophical treatments of information that try to redress this neglect are Mann (2018) and Rathkopf (2017).

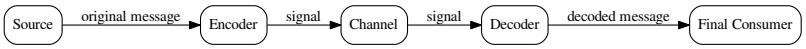


Figure 1: An information-processing pipeline

Here an *encoder* produces a signal as a response to information incoming from a source. This signal goes through a channel and is subsequently decoded, producing a message that is then utilized for whatever purposes downstream. The first thing to note is that the broadly Dretskean ideas about the content of a signal introduced in section 1 only have use for the first two links in this information-processing chain: how signals carry information about a certain original message produced by a source, as depicted in fig. 2. In fact, in information theory the main action happens immediately after that: a source is producing stuff, and we want that stuff to *go through a channel*. Information theory is mainly about providing theoretical guarantees of faithfulness in transmission, given the rate of the channel. We can think of this rate as the number of bits it provides for the encoder to use in the signal. If, say, the rate is 2 bits per use of the channel, this means the encoder can use up to 2 bits to construct the signal and be sure that it can pass unscathed through the channel and on to the decoder.

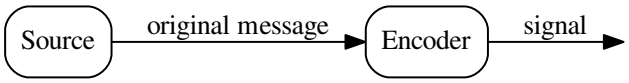


Figure 2: The information-processing pipeline in the Dretskean tradition

In typical cases of representation, channel rate is consistently smaller than ideal. Consider animal alarm calls. Vervet monkeys, for example, are typically described as being able to produce three different, discrete kinds of calls (Seyfarth, Cheney & Marler 1980a, 1980b) that are usually taken to be associated with the presence of leopards, eagles and snakes respectively. Obviously, the entropy of the relevant aspects of the environment that prompt the production of a call (think of all the possible patterns of approach of these predators, for example) vastly outstrip the rate of a channel, which consists in the production of just one out of three possible signals. This means that loss in communication is inevitable. Alarm calls, and for analogous reasons representations in general, are all about *lossy transmission*.

The way in which information theory deals with lossy transmission is by defining a *distortion measure* (Cover & Thomas 2006, p. 304) that gives a score to a pair composed of a certain original message M, and the decoded version thereof, $\hat{M}$. In what follows I

will be using the *Hamming distortion* which simply adds 1 to the distortion when the bits in the original and decoded signals (which we can assume to be binary strings) do not coincide, and 0 otherwise, then normalizes. So, for example, the Hamming distortion between an original signal $M = 010011$ and a decoded signal $\hat{M} = 100010$ is ⅓, because the first, second, and last (a total of 3) bits have been decoded incorrectly, and there are 6 bits in total.

The central result in this so-called *rate-distortion theory* approach to lossy transmission is that there is a *rate-distortion function*, $R(D)$, which gives the minimum rate at which any given distortion is achievable. The actual mathematical expression of the rate-distortion function need not detain us here (see Cover & Thomas 2006, p. 307, theorem 10.2.1), but it is such that the *Blahut-Arimoto* algorithm (Blahut 1972; Arimoto 1972) allows us to calculate it easily.

The main thesis of this paper is that representations belong in information-processing pipelines whose rate-distortion function has *sweet spots*: by this I mean points in the rate-distortion curve such that the usefulness of increasing the rate of the channel past those points is much smaller than before reaching them. Moreover, the encoding-decoding strategies that make use of these representations tend to live in the vicinity of those sweet spots. I submit that it is these information-theoretic properties that the conditions on representation discussed in section 2 try to get at.

To see how rate-distortion analyses work let's start by looking into a source that models a series of fair-coin tosses: this random variable would have two values, *heads* and *tails*, with associated probabilities $P_{heads} = P_{tails} = .5$). Using the Hamming distortion as our target distortion measure, if the coin lands heads (tails) and the decoded message is tails (heads) the distortion is 1, otherwise 0. The Blahut-Arimoto algorithm allows us to draw the rate-disortion curve, in fig. 3. Here the blue line is the rate-distortion curve. It intersects the x-axis at 1.0 bits (the entropy of the source) and it intersects the y-axis at 0.5 (the lowest average distortion one can achieve when the channel is closed.) The red line gives a measure of how steep the blue line is at any given point—in particular, the absolute value of the slope of the blue line. The higher the red line, the steeper the blue line.

The situation this setup is modeling is one in which a single cue is present or absent, and a signal tries to keep track of whether it does. This is precisely the kind of situation where many theorists (certainly Sterelny and Burge, for the reasons reviewed in 2.1) would see the postulation of representations as entirely idle—see, e.g., Schulte's vasopressin example in his Schulte (2015). In agreement with the idea that postulating representations here is idle, there is not much structure to the rate-distortion curve corresponding to this setup: reading the chart from right to left, increasing the rate makes the achievable expected distortion go smoothly down, until the rate hits the entropy of the source, at which point the achievable distortion is zero.That's about it.

Let's now model one kind of situation in which there is a reasonably wide consensus that representations make an explanatory contribution: vervet-monkey alarm calls, as reviewed above. In the model, the source—the situation the information-processing pipeline is dealing with—randomly makes members of two natural kinds (we can think

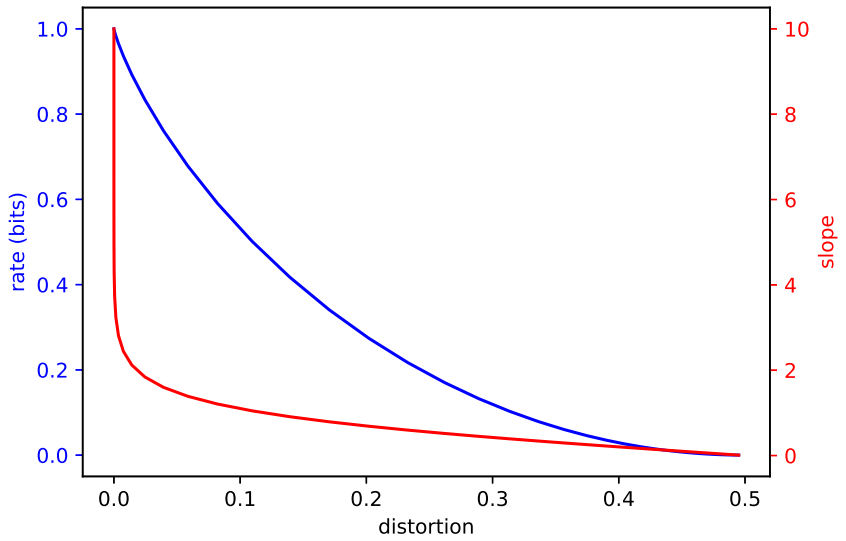


Figure 3: The rate-distortion function for a coin toss

of them as two different predators) be or not present at any given time, independently from one another. This intends to mimic the situation vervet monkeys face, where snakes, leopards and eagles show up or not, more or less at random.

These natural kinds are modeled as homeostatic property clusters (see section 2.2 above). In order to derive a explicit probability distribution for the source out of this qualitative description, the two HPCs are in their turn represented by two Bayesian networks, each with a parent node and four children (see fig. 4.) Each of the nodes stands for a property; if the node is *on* it means the corresponding property is instantiated; if it is *off* it means it is not. In the model, children nodes replicate noisily the state of their parent. Thus, e.g., if the parent is *on* (if the corresponding property is instantiated) each child property will have a .95 chance of being instantiated too; if the parent is *off* the probability for each children of being instantiated is .05. The unconditional probability of instantiation for the two parent nodes is .5.

In the model, the source produces a binary string, with each member of the string being 1 if the corresponding node is on, and 0 if it's off. This signal is encoded, goes through a channel, and is then decoded at the other side. The target distortion measure is the Hamming distortion. Fig. 5 plots the rate-distortion curve for this model.

This curve is very different from the one in fig. 3: there is a clear "sweet spot"—a sudden drop in the usefulness of extra rate, see the red curve—when the system hits a rate of 2 bit/use. I.e, there is, in a certain principled sense, an optimal level of lossy compression; a way to set up an encoding-decoding strategy that recover most of what's going on in

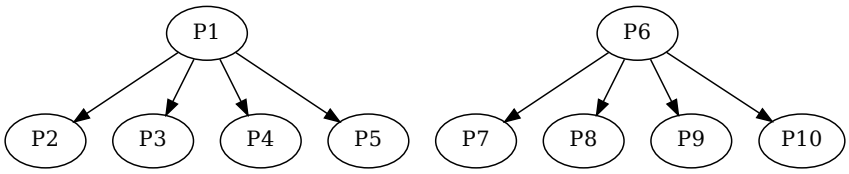


Figure 4: Two natural kinds

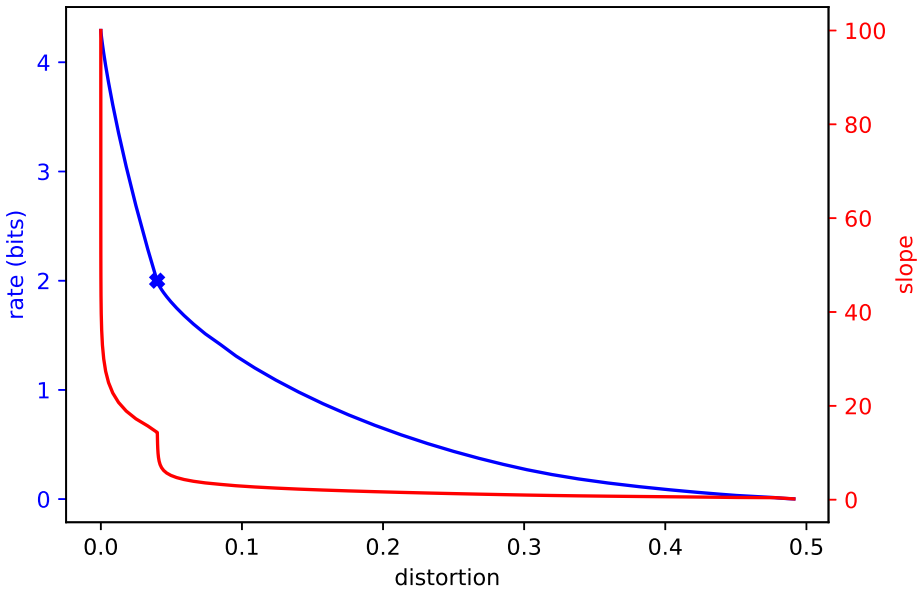


Figure 5: A sweet spot in the rate-distortion function

the world of relevance to the information-processing system, even through a very severe, 2 bit bottleneck. I claim that this is no coincidence. Our representation-attributing practices gravitate towards this kind of situations.

To see how sweet spots in rate-distortion curves and representations are related, consider now what an optimal encoding-decoding strategy would look like. That is, how should the encoder encode the information coming from the source, and how should the decoder decode the signal coming from the encoder, so that the resulting expected distortion between original and decoded signal is the minimum achievable, at the sweet spot?

Optimal Encoding Strategy: First divide the incoming signal in two halves, one corresponding to properties P_1 through P_5; the other corresponding to properties P_6 through P_{10}.

If there is a majority of 1s in the first half of the original signal set the first bit of the signal to 1. Otherwise set it to 0. Ditto for the second half of the original signal and the second bit of the signal.

Optimal Decoding Strategy: If the first bit in the incoming signal is 1, set the first half of the decoded signal to 11111. Otherwise, set it to 00000. Ditto for the second bit and the second half of the decoded signal.

How should we interpret what encoder and decoder are doing here? A natural way is this: they are using the presence or absence of properties in an HPC cluster as diagnostic of the presence or absence of the underlying natural kind—this would be the encoding part—and then taking the resulting signals as representing the presence of a paradigmatic instance of the kind, one that has all the properties in the cluster—this would be the decoding part. HPC kinds being what they are, frequently the first half of the incoming signal will resemble the paradigmatic presence of the first kind (11111) or its paradigmatic absence (00000), and the same will happen with the second half and the second kind. That is why this encoding-decoding strategy works so well.

In describing this optimal strategy I have helped myself to representational vocabulary; it has been useful in order to explain how the strategy works, and how come that behaving in this particular way achieves low distortion at low rates: it is because each of the two bits in the signal is caused by, and causes, behavior that is optimally attuned to the probabilistic structure of each of the two natural kinds in the model world, respectively. Nothing going on in this system falls outside the purview of Shannonian information theory—of information theory *tout court*, so at least in this kind of cases representational talk depends on no non-informational fact.

We can now understand better what's lacking in the philosopher of mind's information-theoretic tookit: it is entirely possible, and computationally trivial, to calculate, e.g., Skyrms's pmi between each of the possible signals (00, 01, 10 and 11) and each of the possible world states (all 1024 of them, from 0000000000 to 1111111111). Doing so would leave us with 4 vectors (one for each signal) with 1024 entries each (one for each world state.) First, this is an unwieldy collection of numbers, which doesn't bring out the relevant structure. For example, if the probability of children nodes being *on* conditional on their parent being *on* was .96 instead of .95 the rate-distortion curve

would be qualitatively identical, with a sweet spot in exactly the same place, yet most numbers in the Skyrmsian informational content vectors would change. Second, and most important, nothing in those 4096 numbers allows us to infer the presence of a sweet spot. The relevant information is simply not there, depending as it does on a distortion measure which is not used in computing Skyrmsian informational contents.

If this is approximately right, the question about what makes representational talk explanatory is readily answered: saying that a certain vehicle is a representation conveys something quite specific about its informational context. It says that the vehicle is part of an encoding-decoding strategy that exploits a sweet spot in a rate-distortion curve—where the curve is in turn fixed by the probabilistic structure of the world, and the target distortion measure. This, in less technical terms, translates to saying that the vehicle is summarizing *relevant* (this is where the distortion measure comes in) aspects of the current situation in an optimal, if lossy, manner, made possible by *how the world* is (this is where the probabilistic structure of the world comes in.) This explication of the explanatory contribution of representations can be turned into an explicit answer to what makes something a representation—an answer, that is, to what Artiga (2016) calls the metasemantic question.

The Rate-Distortion Approach: A signal, S, in a certain information-processing pipeline, P, is a representation if the following two conditions are met:

Existence: There are sweet spots in the rate-distortion curve associated with P.

Optimality: S is produced as part of an encoder-decoder strategy that occupies the vicinity of one of these sweet spots.

So, *pace* Dretske, the core information-theoretic notions of entropy, rate, distortion, etc. *can* provide invaluable insight into the representational status of individual signals. If the rate-distortion approach is on the right track, those information-theoretic notions, through the existence condition, specify the kind of setup where representations live, which then the optimality condition can use to provide a criterion for the representational status of individual signals.

I offer the foregoing discussion as a preliminary case for the rate-distortion approach to representation: it shows how postulating representations is explanatory, even if these representations depend just on (Shannon) information. It illuminates the difference in representational status between cue-driven examples, such as Schulte's vasopressin; and vervet alarm calls, and other similar examples. To complete my case I now show how the ways to bridge the gap between natural and nonnatural information discussed in section 2 can be seen as unwitting attempts to get at rate-distortion sweet spots.

4 There is no Gap to Bridge

What does it take for the existence condition to be met? That is to say, what circumstances result in sudden drops in the slope of the rate-distortion curve? We have seen one such family of circumstances: if the pattern in which properties are instantiated

in the source is noisily replicated in a cluster then sudden drops are to be expected: distortion will decrease with rate up to the point where all the main sources of variation in property instantiations are accounted for, and all that remains is the residual noise in instantiations within each cluster. Take a look again at figs. 4 and 5: to describe this source we basically need enough rate to account for the two main sources of variation: P_1 and P_6. This is not all there is to the world, because it's possible for the other properties to (fail to) token independently of their parent, but the unlikeliness of these departures makes the extra rate comparatively less useful.

Noisy replication of property instantiations is at the core of the HPC theory of natural kinds, as we saw above. This means that, in general, the presence of HPC natural kinds in a source will create sweet spots. This opens a line of argument in favor of reference magnetism from information-theoretic premises: reference magneticsm should be seen as making a point about the kind of probabilistic structure that an information-processing pipeline must be attuned to, if signals are to effect the kind of optimal lossy compression that underlies our representation-attributing practices. Reference magnetism is just a way of meeting part of the existence condition.

Regarding the suggestion, by Sterelny, Burge and others, that representations inhere preferably on signals sitting in a one-to-many-to-one pipeline, I submit that the many-to-one aspect of this suggestion aims at meeting the optimality condition; the one-to-many aspect, together with reference magnetism, aims at meeting the existence condition.

The first thing to note here is that the *Optimal Encoding Strategy* presented above enforces what Sterelny calls robust tracking and Burge calls constancy: the strategy consists in considering all properties coming from each of the two clusters and setting the relevant bit to 1 only if a majority of those properties are instantiated. That is, the encoder is taking a multiplicity of configurations (e.g., the first half of the incoming signal being 00111, 01011, 10111, etc.) to a single output: the first bit of the signal being 1. Furthermore, that part of the signal will be decoded as 11111: from there on, the system downstream will treat whatever is out there in the world as a paradigmatic member of the first kind. The system is recovering the presence of a natural kind out of many different, noisy instantiation patterns. This is a clear instance of constancy. Suppose that the encoder, insted of being many-to-one, depended on a single cue; say, suppose it set the first bit to 1 if one of the children properties (say, P_2) was instantiated, and to 0 otherwise. In such a cue-driven setup, the best encoder-decoder arrangement possible is marked by the blue circle in fig. 6. This has double the distortion than than the optimal encoding (marked by the blue cross) which sits right on top of the optimal rate-distortion curve. This cue-driven system would not meet the optimality condition, which means that a many-to-one architecture is instrumental to meeting it.

Finally, the target distortion measure in the information-processing pipeline can be seen as that which Sterelny's one-to-many condition on representation is actually tracking. Using, for example, the Hamming distance as a distortion measure is tantamount to assuming that all of the properties of the natural kinds are relevant for downstream processing. One natural way in which this may happen is when the agent is to respond flexibly to the presence of the natural kind: in different contexts or states different properties of the kind might be relevant and, for example, the presence of a tree might

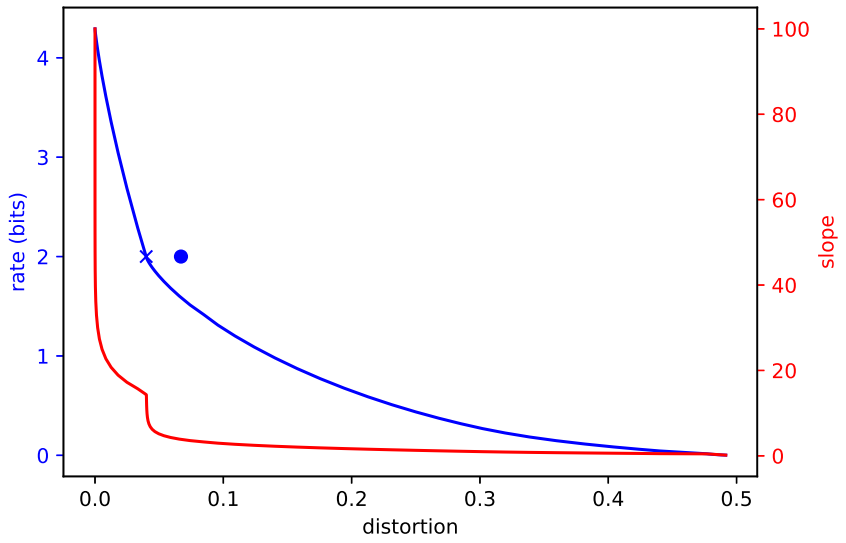


Figure 6: Cue-driven encoding

be sometimes relevant to behavior because it bears fruit (if the agent is hungry) and some other times because it has a dense cover (if the agent is looking for shelter.)

Caring about all (or many) properties of the kind is what makes the rate-distortion curve display a sweet spot. If, instead, the agent has a rigid, stereotyped response to the presence of members of the kinds—that is, if it only cares about the presence of one property, which is the property that makes that rigid behavioral response fitness-conducive, then the curve is as presented in fig. 7. Rigid behavioral responses make the probabilistic structure of the kinds largely irrelevant. As a result, the system behaves as if a coin were tossed, where heads would mean that the target property is tokened, and tails that it is not. This arrangement does not meet the existence condition. Sterelny's broad-banded responses are, again, a way of getting at rate-distortion sweet spots.

References

Arimoto, S 1972, 'An algorithm for computing the capacity of arbitrary discrete memoryless channels', *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20.

Artiga, M forthcoming, 'Beyond Black Spots and Nutritious Things: A Solution to the Indeterminacy Problem', *Dialectica*.

Artiga, M 2016, 'Liberal Representationalism: A Deflationist Defense', *dialectica*, vol.

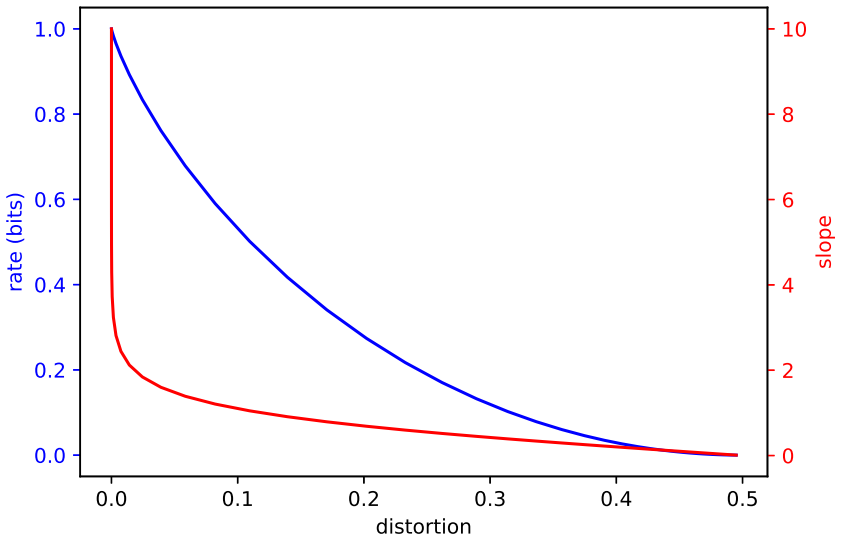


Figure 7: Rigid behavioral response

70, no. 3, pp. 407–430.

Blahut, R 1972, 'Computation of channel capacity and rate-distortion functions', *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473.

Boyd, R 1989, 'What Realism Implies and What It Does Not', *Dialectica*, vol. 43, no. 1-2, pp. 5–29.

Burge, T 2010, *Origins of objectivity*, Oxford University Press.

Cover, TM & Thomas, JA 2006, *Elements of Information Theory*, New York: Wiley.

Dretske, F 1981, *Knowledge and the Flow of Information*, The MIT Press.

Lewis, D 1983, 'New work for a theory of universals', *Australasian journal of Philosophy*, vol. 61, no. 4, pp. 343–377.

Lewis, D 1984, 'Putnam's paradox', *Australasian Journal of Philosophy*, vol. 62, no. 3, pp. 221–236.

Mann, SF 2018, 'Consequences of a Functional Account of Information', *Review of Philosophy and Psychology*, pp. 1–19.

Martínez, M 2013, 'Teleosemantics and Indeterminacy', *Dialectica*, vol. 67, no. 4, pp.

427–453.

Millikan, R 1984, *Language, Thought and Other Biological Categories*, The MIT Press.

Papineau, D 1987, *Reality and Representation*, Basil Blackwell.

Rathkopf, C 2017, 'Neural information and the problem of objectivity', *Biology & Philosophy*, vol. 32, no. 3, pp. 321–336.

Ryder, D 2006, 'On Thinking of Kinds', in G Macdonald & D Papineau (eds), *Teleosemantics*, Oxford University Press, pp. 1–22.

Ryder, D 2004, 'SINBAD Neurosemantics: A Theory of Mental Representation', *Mind & Language*, vol. 19, no. 2, pp. 211–240.

Scarantino, A 2015, 'Information as a probabilistic difference maker', *Australasian Journal of Philosophy*, vol. 93, no. 3, pp. 419–443.

Schulte, P 2015, 'Perceptual representations: A teleosemantic answer to the breadth-of-application problem', *Biology & Philosophy*, vol. 30, no. 1, pp. 119–136.

Seyfarth, RM, Cheney, DL & Marler, P 1980a, 'Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication', *Science*, vol. 210, no. 4471, pp. 801–803.

Seyfarth, RM, Cheney, DL & Marler, P 1980b, 'Vervet monkey alarm calls: Semantic communication in a free-ranging primate', *Animal Behaviour*, vol. 28, no. 4, pp. 1070–1094.

Sider, T 2014, *Writing the Book of the World*, Reprint edition., Oxford University Press, Oxford.

Skyrms, B 2010, *Signals: Evolution, Learning & Information*, New York: Oxford University Press.

Sterelny, K 2003, *Thought In A Hostile World: The Evolution of Human Cognition*, John Wiley & Sons, Malden, MA.

The Proportionality of Common Sense Causal Claims

Jennifer McDonald

This paper defends strong proportionality against what I take to be its principal objection – that proportionality fails to preserve common sense causal intuitions – by articulating independently plausible constraints on representing causal situations. I first assume the interventionist formulation of proportionality, following Woodward.[1] This views proportionality as a relational constraint on variable selection in causal modeling that requires that changes in the cause variable line up with those in the effect variable. I then argue that the principal objection derives from a failure to recognize two constraints on variable selection presupposed by interventionism: *exhaustivity* and *exclusivity*.

[1] Woodward 2003

1. Introduction

Yablo's principle of proportionality holds, roughly, that something counts as a cause of some effect just in case it includes the appropriate degree of causal information.[2] Proportionality has been put to various philosophical uses, such as a proposed solution for the causal exclusion argument, and as a justification and explanation of the dependence on high-level causal explanations in the special sciences. However, the precise formulation of such a principle has proven to be controversial.

I take the most promising formulation to be an interventionist one, following Woodward.[3] Such a formulation defines proportionality as a relational constraint on variable selection in causal modeling. In this paper, I argue that this formulation works well as it is – contra Franklin-Hall (see 2016) – so long as we recognize two independently plausible background requirements on variable selection. I call these *exhaustivity* and *exclusivity*. Exhaustivity holds that a variable must take at least one of its values. Exclusivity holds that a variable can take at most one of its values. Both constraints are relative to, and thereby help to make explicit, the modal assumptions implicit in causal inquiry.

Finally, with these requirements in place, I defend proportionality against its principal objection: that it fails to preserve fundamental causal intuitions. I demonstrate how this concern derives from a failure to recognize and integrate the modal assumptions implicit in causal inquiry, in tandem with an inappropriate use of variables to represent causal situations.

2. Interventionism

The formulation of proportionality that I endorse comes directly from Woodward, and is defined in terms of his interventionist account of causation. Interventionism expands on the intuition that causal claims provide

[2] Yablo 1992
[3] Woodward 2003, 2008a, 2008b, 2010, 2016

manipulability information. If *X* causes *Y*, then manipulating or changing *X* is a way of manipulating *Y*. It then exploits the language of causal models to identify and articulate different causal relations of interest. A causal model can take a variety of forms, such as graphical, potential-outcome, and structural-equations models.[4] However, I'll restrict discussion of causal models in this paper to graphical models. A graphical model is, essentially, a set of variables – representing the causal relata – and a directed binary relation between them – representing causal influence.

Interventionism then defines the notion of an *intervention* on a system. An intervention, *I*, first must directly change the value of some variable, *X*, in such a way that it breaks the dependence that *X* may have had on other variables in the system. Second, *I* must be designed in such a way that any change in the effect variable, *Y*, will be the direct result of *X* and not of *I* itself. Finally, *I* must be wholly independent of other possible causes of *Y*, whether such causes are represented by the given model or not. A more precise formulation than this won't matter for the purposes of this paper.[5]

With this in place, the interventionist then defines a basic notion of cause, which corresponds most closely with the intuitive notion of *causal relevance*:

> (Principle M) *X* causes *Y* iff there are background circumstances B such that if some (single) intervention that changes the value of *X* (and no other variable) were to occur in B, then Y would change. (Woodward 2003, 222)

That is, in order for *X* to be a cause of *Y*, the change in *X* from one value to another as the result of an intervention corresponds to the change in *Y* from one value to another, given some fixed set of background parameters. Various kinds of causal relations are then captured by refinements on this basic notion. Due to

[4] See Greenland and Brumback 2002 and Hitchcock 2009 for overviews of causal models.
[5] See Woodward 2003, chapter 3, especially 98

the irrelevance of these and further details to my argument, I'll leave my overview of interventionism here.[6]

3. Proportionality as Relational Constraint on Variable Selection

Interventionism places variables front and center in how we represent and inquire into causation. Thus, more needs to be said about the criteria for variable selection. Although the variables can be taken to represent different things, I will assume throughout that the set of values of a particular variable represents a set of properties – constrained by a given property type – that are possibly instantiated by some particular thing. The assumed causal relata of this paper will therefore be property instantiations.

This paper addresses two questions relevant to variable selection: (i) What determines the range of values that a variable can take? (ii) At what level of description should the values of the variables be? Proportionality has been proposed as an answer to (ii). However, after laying out the proposal, I'll go on to argue that while (ii) can be answered by the principle of proportionality, it can only do so alongside an appropriate answer to (i). One aspect of such an answer is that the background modal context determines the range of values that a variable takes.

Constraints on variable selection can be divided into two kinds: relational constraints and non-relational constraints. *Relational constraints* pertain to the extrinsic nature of the variables in a causal model, to how "variables relate to one another." (Woodward 2016, 1056) One example of such a constraint is stability.[7] *Stability* is the persistence of the causal relation between a cause variable and an effect variable, despite changes in the background conditions. The more changes such a relation can survive, the more stable it is.

[6] See Woodward 2003, chapter 2, especially section 3
[7] See Woodward 2010, 2016

Proportionality is just such a relational constraint. It holds that changes in a cause variable should line up with changes in an effect variable. Intuitively,

> Proportionality has to do with whether changes in the state of the cause 'line up' in the right way with changes in the state of the effect and with whether the cause and effect are characterized in a way that contains irrelevant detail. (Woodward 2010, 287)

Take Yablo's pigeon example.[8] Sophie the pigeon is trained to peck at red things and only at red things. She then pecks at a paint chip, which is a particular shade of red – scarlet. Which of the following is causally relevant to Sophie's pecking: the chip's being red or the chip's being scarlet?

When translated into interventionist terms, this becomes a false dichotomy. Take the variable, *P*, to be a variable representing whether the pigeon pecks or not. It can take the values: {*peck*, *not-peck*}. Now consider two alternative variables for representing the property-instantiations of the paint chip: the variable, *R*, which can take the values {*red*, *not-red*}, and the variable, *T*, which can take the values {*taupe, scarlet, cyan, mauve, crimson,* etc.}, where 'etc.' stands for all other physically possible colors at the same grain as those already made explicit. According to Principle M, the causal model in which *R* stands as causally relevant to *P* is just as accurate as one in which *T* so stands. In the *R* model, *R* is causally relevant to *P* because an intervention on *R* that changes its value from *not-red* to *red* changes *P*'s value from *not-peck* to *peck.* In the *T* model, *T* is causally relevant to *P* because an intervention on *T* that changes its value from *taupe* to *scarlet* changes *P*'s value from *not-peck* to *peck*.

Interventionism therefore doesn't ask the question, which variable stands in a causal relation to *P*? For, the answer is 'both'. *R* and *T* are each causally relevant to *P*. But, this doesn't mean that their respective relationship to *P* is the same. *R* is *proportional* to *P*, while *T* is not. All of the changes in *R* line up with changes in *P* – every intervention on *R* corresponds to a change in *P*. But only some of the

[8] Yablo 1992

changes in *T* line up with those in *P* – only certain interventions on *T* correspond to changes in *P*. The intervention that changes the value of *T* from *taupe* to *cyan*, for example, will not change the value of *P*.

Woodward defines proportionality more explicitly as,

> (P) There is a pattern of systematic counterfactual dependence (with the dependence understood along interventionist lines) between different possible states of the cause and the different possible states of the effect, where this pattern of dependence at least approximates to the following ideal: [it] should be such that (a) it explicitly or implicitly conveys accurate information about the conditions under which alternative states of the effect will be realized and (b) it conveys only such information – that is, the cause is not characterized in such a way that alternative states of it fail to be associated with changes in the effect. (2010, 298)

There are two views on what this difference between variables like *R* and *T* means. The first takes proportional variables such as *R* to represent genuine causes, while non-proportional variables such as *T* represent merely causally relevant factors. Proportionality is thereby considered a necessary constraint on causation. Call this *strong proportionality*.[9] The second view takes proportionality to be a merely pragmatic constraint on causal explanation.[10] Call this *weak proportionality*. Throughout this paper, I assume and defend strong proportionality.

4. Non-Relational Constraints: Exhaustivity and Exclusivity

Non-relational constraints, on the other hand, pertain to the intrinsic nature of the variables in a causal model. These constraints "can be applied to variables, individually, independently of how they relate to other variables." (Woodward

[9] See List and Menzies 2009; Menzies and List 2010; and Papineau 2013
[10] See Woodward 2015; Shapiro and Sober 2012; McDonnell 2017; and Weslake 2013, 2017

2016, 1057) One example is *metaphysical naturalness*, which requires that variables pick out only natural properties, on some understanding of 'natural'.[11]

What I propose to call the exhaustivity and the exclusivity constraint are similarly non-relational constraints. Take exhaustivity first. The *exhaustivity constraint* requires that a variable's values capture the entire range of relevant possibilities for whatever type of thing the variable represents. An exhaustive variable is one that must take one of its values, given whatever background modal constraints are in place.

Since I've restricted this discussion to variables whose values represent the property instantiation of some target object, I can define exhaustivity in more precise terms. *Exhaustivity* is the constraint on a variable in a causal model that holds that its values must jointly represent the range of possibilities of property instantiation by the given object for the given property-type. If the property-type is a color, for example, then the values must somehow exhaust the color spectrum. This can be done quite simply with a binary variable that can take the values: {*some particular color*, *not-(that particular color)*}.

Next, the *exclusivity constraint* holds that the values of a given variable should be such that any one excludes all the others. Woodward references exclusivity when he writes,

> When considering the values of a single variable, we want those values to be logically exclusive, in the sense that variable *X*'s taking value *v* excludes *X*'s also taking value *v'*, where $v \neq v'$. (2016, 1064)

In other words, if two things are not exclusive – if they could occur together – then they should be represented by distinct variables. While exhaustivity holds that a variable should take *at least* one of its values, exclusivity holds that a variable should take *at most* one of its values.

[11] See Lewis 1983; Menzies 1996; Paul 2000; and Franklin-Hall 2016

Importantly, exhaustivity and exclusivity are each relative to a background modal context. In possible worlds terminology, the modal context is the set of possible worlds relevant to the truth of the counterfactual that captures the causal claim. It can be described as a set of worlds, or perhaps more succinctly as a list of background assumptions that define such a set. These assumptions can include any constraint that operates in a law-like fashion.

For example, the causal claim, "The chip's being scarlet caused the pigeon to peck," corresponds to the counterfactual, "Had the chip not been scarlet, the pigeon wouldn't have pecked." The modal context of this claim and corresponding counterfactual is the set of possible worlds that determines whether the counterfactual is true. So, if this claim and counterfactual are meant to represent a *specific* causal situation near a local paint chip factory that specializes in just the colors scarlet and cyan, and no others, then the relevant set of possible worlds will be constrained to those in which the paint chip takes one of the two factory colors – cyan or scarlet. In this context, the variable, *C,* that can take the values {*cyan, scarlet*}, is an exhaustive variable. Further, given this set of worlds, the counterfactual is true.

If instead these are meant to represent any *general* causal situation involving paint chips and a red-pecking pigeon, then the relevant set of possible worlds will be more inclusive, including all worlds in which the paint chip takes any color within the color spectrum. *C* is not exhaustive relative to this more inclusive modal context. But the variable *T*, from before, is. Given this more inclusive set of worlds, the counterfactual is false, since the pigeon will peck in response to shades of red other than scarlet.

A point of note here is that the constraints of exhaustivity and exclusivity are indeed non-relational constraints in the sense previously defined. Although they are relative to the modal context, they are *not* relative to other variables in the model. They are properties of a variable taken independently as a representation of the target scenario.

I hold that causal models successfully represent causal situations in part by requiring exhaustive and exclusive variables. Proportionality, defined in terms of causal models, also requires exhaustive and exclusive variables. A significant upshot of this is that the proportional cause is not only relative to the target effect variable, but also to the background modal context.

5. Interventionist Proportionality Does the Trick

Franklin-Hall contends that Woodward's formulation of proportionality doesn't successfully prioritize intuitively proportional causal relata, such as red in the pigeon example. However, as I'll argue, presupposing my notion of exhaustivity corrects for this objection.

Franklin-Hall argues that proportionality as laid out in section 3 is inadequate for capturing the kind of causal explanation we're looking for. To do so, she calls upon Sophie and her paint chip. She then introduces a comparison between the causal variable, *R*, that can take the values: {*red*, *not-red*}, (as above), and a variable, *C*, that can instead take the values: {*cyan, scarlet*} (as above). *R*, as before, is proportional to, and therefore a genuine cause of, *Y*. But, she argues, *C*, too, is proportional to *Y*, since every possible intervention on *C* changes the value of *Y*. An intervention on *C* that changes its value from *cyan* to *scarlet* changes *Y* from *not-peck* to *peck*, and an intervention that changes *C*'s value from *scarlet* to *cyan* changes *Y*'s value from *peck* to *not-peck*. Thus, the changes in *C* line up with the changes in *Y* just as well as the changes in *R* do. The problem, then, is that proportionality, as formulated, is insufficient to its intended task. It fails to privilege a variable like *R* over one like *C*, and so fails to prioritize a causal model that uses *R* over one that uses *C*.

In response to this problem, a natural move would be to find a way to disqualify variables like *C* from the arena. Intuitively, *C* is not the right kind of variable. But, why not? I propose that our aversion to variables like *C* is due to their failure to exhaustively represent the implicit modal context of the situation. The background possibilities relative to the paint chip include the full color spectrum.

Unless the possible color of the paint chip is restricted in some way – by the local factory, for example – then the target object can fail to take one of *C*'s two values. There are other physically possible colors that the paint chip could have – such as beige or olive green – and *C*'s values fail to represent these possibilities.

Relative to the implicit modal context, then, *C* is not an exhaustive variable. The variable, *R*, on the other hand, is exhaustive, since the object must take one of *R*'s two values. By requiring exhaustive variables, *C* is discounted as a candidate variable *relative to the implicit modal context*, and *R* takes privilege as the proportional cause.

In general, two variables are in proper competition with each other over which is proportional to some effect variable only when they are exhaustive relative to the same modal context. *C* and *R* are not competitors for proportionality relative to *Y*, since only one of them can contain an exhaustive set of active possibilities relative to any given modal context.

6. Preserving Causal Intuitions

The strongest objection to proportionality, as raised by Bontly, Shapiro and Sober, McDonnell, and Weslake, is that it seems to render many common sense causal claims false.[12] Call this the *objection from common sense*. It objects to strong proportionality by attempting to demonstrate that if proportionality is required of something to be a cause, then many things that we would naturally call causes don't actually qualify.

Take as an example the situation where Socrates drinks hemlock and then dies, and the corresponding causal claim, 'Socrates's drinking hemlock caused him to die'. The objection goes that drinking hemlock is not actually proportional to Socrates dying. For example, if Socrates had not drank hemlock, but still consumed it – by eating a dozen leaves, for example – then he still would have

[12] See Bontly 2005; Shapiro and Sober 2012; McDonnell 2017; and Weslake 2013, 2017

died. This seems to show that the changes in the variable that represents Socrates drinking hemlock don't line up with the changes in the variable that represents Socrates dying. The first variable could change values from *Socrates-drinks-hemlock* to *Socrates-eats-hemlock* and the second variable would retain the value *Socrates-dies*. This common sense causal claim is therefore not proportional. The proportional cause should be, instead, *consuming hemlock*.

However, this objection is mistaken. It fails to respect the exhaustivity constraint on variable selection, and thereby equivocates between different background modal contexts. It further fails to respect exclusivity, and thereby runs together what should be different variables. Rectifying this illuminates the implicit proportionality of common sense causal claims.

First, the objection ignores the fact that proportionality, in requiring exhaustive and exclusive variables, is relative to modal context. Take the hemlock example just outlined. Importantly, this example and corresponding claim are under-defined.[13] Translated into interventionist terms, all that this description provides is that there is some variable that takes a value that represents Socrates drinking hemlock, and an intervention on this variable changes the value of some other variable to one that represents Socrates dying. But, a number of different variables could represent the purported cause, and a number of different models could represent its relationship to the effect of Socrates' dying. Which of these is accurate depends on what the relevant alternatives to drinking hemlock are. How these details get filled in will determine whether or not the variable that represents Socrates drinking hemlock is proportional.

I hold that the common sense claim that drinking hemlock causes Socrates's death implicitly takes the relevant alternative to be Socrates's *not* drinking hemlock. The default context is taken to be that hemlock was the only possible poison, and drinking it the only possible means of consumption. Given this context, the exhaustive variable would take the values {*drinks-hemlock*, *doesn't-*

[13] I take this to be common knowledge. See Franklin-Hall 2016; McDonnell 2017; and Weslake 2017

drink-hemlock}. But, such a variable is indeed proportional to the effect variable. Thus, the common sense cause is, in fact, proportional.

Such a defense requires that common sense claims be implicitly relative to a modal context. I'm not the first to relativize common sense claims to context. Philosophers such as Mackie and Schaffer make such a move, albeit with different ends in mind.[14] However, both McDonnell and Weslake explicitly deny this kind of relativity.[15] They claim that the very fact that we have strong and convergent intuitions about common sense examples, despite their being under-determined, demonstrates that the intuitions are not sensitive to filling in details.

In response, I argue that we respond to common sense causal examples in the same way that we respond to standard conversations. According to Grice, communication is governed by a set of conversational maxims.[16, 17] The maxims most relevant to how an audience engages with these under-defined causal examples are the maxims of *quantity* and *relation*. Taken together, these maxims enjoin an interlocutor to,

> Make your contribution as informative as is required (for the current purposes of exchange)….[and no] more informative than is required,….[and b]e relevant. (1989, 26 – 27)

Thus, the conversationally natural way to fill in the modal context of these examples is to take each fact as informative and relevant, and to assume that all informative facts have been provided.

The only information provided by the hemlock example is the following: (i) Socrates drinks hemlock. (ii) Socrates dies. The Gricean maxims tell us that this is all the information needed, and that nothing significant has been left out. So, the details are filled in as continuous with everyday life. In possible world speak,

[14] See Mackie 1974, especially chapter 2; and Schaffer 2005
[15] McDonnell 2017; Weslake 2017
[16] See Grice 1989
[17] Bontly makes a similar point (see 2005)

we're looking only at worlds which have a similar environment, a biologically similar Socrates, etc., and in which laws of metaphysical necessity hold.

The causal focus is on Socrates's drinking hemlock. This means that in evaluating the causal relationship, everything else is held fixed and the fact of the drinking hemlock is varied. Due to the absence of any other details, the only real alternative to Socrates's drinking hemlock is his not drinking hemlock. Nothing suggests that there are alternative means of consuming the hemlock. Further, it's not a common occurrence in everyday life to have alternative means of consuming a given poison. Treating *eating hemlock* as a relevant alternative would be to arbitrarily introduce something that wasn't otherwise specified, and whose presence can't be justified by everyday experience.

The objection from common sense assumes different possible alternatives than what I take to be implicit, and then tries to say that relative to these other alternatives, the common sense causal claim is not proportional. I have argued that the common sense cause is simply not relative to these other alternatives.

However, even given other possible alternatives, the common sense cause would still be proportional. The second mistake that the objection makes is that it fails to appreciate the constraint of exclusivity.

The objection holds that there is some relevant alternative to Socrates's drinking hemlock that preserves his consuming it. Take as an arbitrary alternative his eating hemlock. Socrates could both drink and eat the hemlock – he could wash down a hemlock salad with a glass of hemlock milk, for example. Following exclusivity, then, these possibilities should be represented by distinct variables – one that can take the value *drinks-hemlock*, call this *D*, and one that can take *eats-hemlock*, call this *E*.

But, now there is no problem. Following Woodward's response to early pre-emption cases,[18] we can hold *E* fixed at the value that represents Socrates not eating the hemlock, and see if the changes in *D* – which we can ensure meets exhaustivity by giving it the second value *doesn't-drink-hemlock* – line up with the changes in the effect variable. They do. When an intervention sets the value of the cause variable to *drinks-hemlock*, the effect variable takes the value *dies*. When an intervention sets the value of the cause variable instead to *doesn't-drink-hemlock*, the effect variable changes value to *doesn't-die*. Once again, the common sense cause is proportional.

If, on the other hand, the situation is such that Socrates's drinking hemlock is indeed mutually exclusive with his eating hemlock, then *drinks-hemlock* and *eats-hemlock* could be values of the same variable. Imagine that Socrates's jailor only has enough money to purchase either hemlock leaves or hemlock milk, but not both. In this case, neither Socrates's drinking nor his eating will be proportional. The proportional cause is instead his consuming hemlock. The proportional variable will therefore be one that takes as values {*consumes-hemlock*, *doesn't-consume-hemlock*}.

But, this is not in conflict with common sense – so long as we abstract away from normal everyday circumstances, and instead genuinely fix the situation as one in which Socrates is forced to consume hemlock, arbitrarily receiving hemlock leaves or milk. When, given this background, we're asked what causes Socrates's death, it is natural to say that it was his consuming hemlock. After all, it isn't the drinking nor the eating that makes a difference to whether Socrates dies, since had he not done one he would have done the other. It is his consuming hemlock rather than not.

Finally, I'd like to point out that the intuition that Socrates's consuming hemlock is the more proportional cause is actually misguided. The naïve intuition holds that an exhaustive and exclusive variable with the value *consumes-hemlock* – call this H_1 – is more proportional to the exhaustive and exclusive variable with the

[18] See Woodward 2003

value *drinks-hemlock* – call this H_2. But, the modal context to which H_1 will be exhaustive is different than that to which H_2 will be. They're therefore not even in competition for proportionality. Instead, I suggest that this intuition is a response to the fact that H_1's modal context is more inclusive than that of H_2. H_1 can accurately (and proportionally) represent the cause of Socrates's death in a wider range of situations than can H_2. But, this is about stability – as earlier defined – not about proportionality. The model that employs H_1 is simply *more stable* than that which employs H_2. This putative proportionality intuition is actually responding to the property of stability.

7. Conclusion

In this paper, I have defended the interventionist formulation of proportionality by explicating the exhaustivity and exclusivity constraints, and stipulating that proportionality requires variables that meet these constraints.

These constraints have been defined on the assumption that a variable represents a particular object's instantiations of a particular type of property. But, they are easily generalized to cover alternate objects of representation. Take events, for example. If variables represent particular kinds of events occurring or failing to occur, then exhaustivity would require that the values of a variable cover the entire range of possibilities of event occurrence for whatever type of event the variable represents. Exclusivity would require that the values of a variable be event occurrences such that no two could occur simultaneously.

Finally, I have articulated how the interventionist formulation of proportionality responds to the objection from common sense. Such an objection dissolves once the explicated constraints on variable selection are honored.

8. References

Bontly, Thomas. 2005. "Proportionality, Causation, and Exclusion." *Philosophia* 32 (1-4): 331 – 48

Franklin-Hall, Laura. 2016. "High-Level Explanation and the Interventionist's 'Variables Problem.'" *British Journal for the Philosophy of Science* 67 (2):553 – 77

Greenland, Sander, and Babette Brumback. 2002. "An Overview of Relations Among Causal Modelling Methods." *International Journal of Epidemiology* 31:1030 – 37

Grice, H. Paul. 1989. *Studies in the Way of Words.* Cambridge, MA: Harvard University Press

Hitchcock, Christopher. 1996. "The Role of Contrast in Causal and Explanatory Claims." *Synthese* 107 (3): 395 – 419

--- 2009. "Causal Modelling." In *The Oxford Handbook of Causation*, ed. Helen Beebee, Christopher Hitchcock, and Peter Menzies, 299 – 314. Oxford: Oxford University Press

Lewis, David. 1983 "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343 – 77

List, Christian, and Peter Menzies. 2009. "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106 (9): 475 – 502

Mackie, J. L. 1974. *The Cement of the Universe.* Oxford: Oxford University Press

McDonnell, Neil. 2017. "Causal Exclusion and the Limits of Proportionality." *Philosophical Studies* 174 (6): 1459 – 74

Menzies, Peter. 1996. "Probabilistic Causation and the Pre-emption Problem." *Mind* 105 (417): 85 – 117

Menzies, Peter, and Christian List. 2010. "The Causal Autonomy of the Special Sciences." In *Emergence in Mind*, ed. Cynthia Macdonald and Graham Macdonald, 108 – 28. Oxford: Oxford University Press

Papineau, David. 2013. "Causation is Macroscopic but not Irreducible." In *Mental Causation and Ontology*, ed. Sophie C. Gibb and Rögnvaldur Ingthorsson, 126 – 52. Oxford: Oxford University Press

Paul, L.A. 2000. "Aspect Causation." *The Journal of Philosophy* 97 (4): 235 – 56

Schaffer, Jonathan. 2005. "Contrastive Causation." *The Philosophical Review* 114 (3): 297 – 328

Shapiro, Larry, and Elliott Sober. 2012. "Against Proportionality." *Analysis* 72 (1): 89 – 93

Weslake, Bradley. 2013. "Proportionality, Contrast, and Explanation." *Australasian Journal of Philosophy* 91 (4): 785 – 97

--- 2017. "Difference-Making, Closure, and Exclusion." In *Making a Difference,* ed. Helen Beebee, Christopher Hitchcock, and Huw Price, 215 – 32. New York: Oxford University Press

Woodward, James. 2003. *Making Things Happen.* Oxford: Oxford University Press

--- 2008a. "Mental Causation and Neural Mechanisms." In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, ed. Jakob Hohwy & Jesper Kallestrup, 218 – 62. Oxford: Oxford University Press

--- 2008b. "Response to Strevens." *Philosophy and Phenomenological Research* 78 (1): 193 – 212

--- 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biological Philosophy* 25 (3): 287 – 318

--- 2015. "Interventionism and Causal Exclusion." *Philosophy and Phenomenological Research* 91 (2): 303 – 47

--- 2016. "The Problem of Variable Choice" *Synthese* 193 (4): 1047 – 72

Yablo, Stephen. 1992. "Mental Causation" The Philosophical Review 101 (2): 245 – 80

Species as Models

Abstract: This paper argues that biological species should be construed as abstract models, rather than biological or even tangible entities. Various (phenetic, cladistic, biological etc.) species concepts are defined as set-theoretic models of formal theories, and their logical connections are illustrated. In this view organisms relate to a species not as instantiations, members, or mereological parts, but rather as phenomena to be represented by the model/species. This sheds new light on the long-standing problems of species and suggests their connection to broader philosophical topics such as model selection, scientific representation, and scientific realism.

1 Introduction

Biological species has arguably been one of the most controversial topics in the philosophy of biology. Philosophers and biologists alike have long debated over "correct" concepts of species and their ontological status. The traditional account took species as a category, class, or type instantiated by individual organisms. After the advent of evolutionary theory, the typological concept came under fire by those who identify species with a part of biological lineage (Ghiselin 1974; Hull 1976). They forcefully

argued that a species is not an abstract type but a concrete historical entity of which individual organisms are mereological bits. Although this individualist thesis became a de-facto standard in the philosophy of biology in the last century, some have complained its lack of explanatory power and called for a revival of a type or natural-kind based concept of biological species (Boyd 1999).

To this debate between individualists and typologists, this paper introduces yet another thesis according to which species taxa are models of scientific theory. Model is a notoriously equivocal concept, but in this paper it is understood as a set-theoretic entity that makes sentences of a given theory true or false. This implies that biological species are mathematical, rather than biological or even tangible, entities. To work out this claim I begin Section 2 with a reconstruction of various (e.g., phenetic, cladistic, biological etc.) species concepts in terms of formal models that licence characteristic sets of inferences. The model-theoretic rendering illustrates logical connections among different species concepts and provides a platform to evaluate them as a problem of *model selection*. Section 3 then expounds on philosophical implications of the model-theoretic interpretation. Identifying species with models entails that the organism-species relationship is not instantial or mereological, but rather representational; i.e., species as models *represent* individual organisms. This opens the possibility of applying general philosophical discussions on scientific representation and realism to vexed questions concerning the epistemic and ontological status of biological species. Through these arguments this paper puts the species problem under broader contexts of model selection, scientific representation, and scientific realism, depicting it as a special case of the generic question as to how science investigates the world.

2 Species as models

This section fleshes out the main claim of this paper by reconstructing various species concepts as set-theoretic models. The central idea is that species concepts specify theories that underpin biological inferences and descriptions, and species are models that satisfy such theories.

2.1 Typological species concepts

The traditional typological view defines species by its essence, or necessary and sufficient conditions or traits. This finds a straightforward expression as a biconditional form $\forall x(Sx \leftrightarrow T_1x \wedge T_2x \wedge \cdots)$. The extension of species S that satisfies this formula then is the intersection $\bigcap_i \mathbf{T}_i$ (see Figure 1(a)).

Though crude as it is, the biconditional formulation allows certain inferences from traits to species and vice versa. It is this kind of logical reasoning that has enabled, for example, the famous French zoologist George Cuvier to reconstruct the anatomy of a whole organism from just a single piece of bone. As is well known, however, such inferences have very restricted validity, because in most cases it is impossible to find a definite set of phenotypic or genetic characteristics that exclusively defines a given species. Evolution implies species boundaries to be necessarily "fuzzy," which undermines simple biconditional forms. The typological species concept has thus been criticized for its lack of expression ability: a simple algebra of trait-sets cannot capture the nuanced reality of biological species.

(a) Typological Sets (b) Cluster Phenotypic space (c) Cladistic Posets (d) Relational Graphs

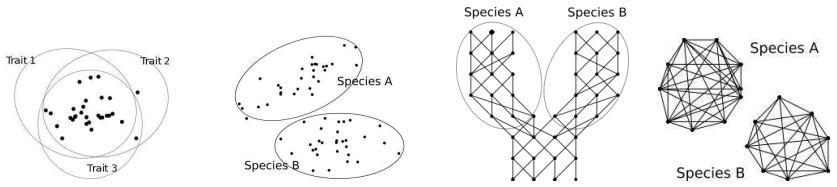


Figure 1: Illustrations of models of various species concepts, with corresponding formal setups. In each model dots/nodes represent individuals. See text for explanation.

2.2 Cluster species concepts

The cluster species concepts avoid this difficulty by defining a species as a group or cluster of similar organisms that do not necessarily share a common set of traits. The question then is how to define similarity. Its earliest variant, the phenetic species concept, represents organisms in a multi-dimensional space each axis of which defines a recorded trait (Sokal and Sneath 1963). Phenotypic similarity is then measured by the euclidean distance between two points/organisms, and a chunk or cluster of organisms in this euclidean space is identified as a species (Figure 1(b)). The choice of euclidean distance is not obligatory. One could, for example, measure similarity by the cosine between two points in the normalized phenotypic space, in which case the similarity amounts to correlation, with a species being identified as a correlated cluster or more generally a *probability distribution* over the phenotypic space (Boyd 1999).

The phenotypic space with a certain metric or probability distribution is certainly a much richer machinery than overlapping sets and allows for more nuanced expressions and inferences. The sophisticated theoretical background (euclidean geometry or

probability theory) enables one to measure the similarity among organisms and to make a trait-species inference in the absence of necessarily or sufficient criteria. To what extent such clustering and inference reflect objective species boundaries, however, was disputed, for the similarity calculation depends much on which phenotypic characters are taken into account. It should also be noted that, like the typological concept, the cluster concepts are purely static and lack a means to express the evolutionary past, the point often criticized by more historical approaches to species.

2.3 Cladistic species concepts

The cladistic species concepts focus on evolutionary history and define species solely in terms of phylogenetic relationships, as a "branch" (monophyletic group) in the evolutionary tree (Hennig 1966). Since ancestral relationship is antisymmetric and transitive, phylogeny forms a (strict) *partially ordered set* or *poset* $(\Omega, \prec)$, with Ω corresponding to a set of organism and $\prec$ meaning "is an ancestor of." A cladistic species is then defined as descendants from some founder organism(s) ω_f:

$$\{\omega \in \Omega : \omega_f \prec \omega\}. \tag{1}$$

An obvious advantage of the cladistic concepts is that it is faithful to the fact of evolution, and for this reason it has been most well received by biologists and philosophers alike. It is not, however, without flaws. For one, although the requirement of monophyly specifies a necessarily condition, it is silent as to how big a branch must be to qualify as a species (for even a small family can satisfy (1)), and so far no satisfactory sufficient condition was given (Velasco 2008). The monophyly requirement has also been

criticized to be too strong, for it would count birds as reptiles because the smallest monophylic group including lizards, snakes, and crocodiles also includes birds. That is, the cladistic species concepts make paraphyletic groups like reptilia *meaningless* (*Sensu* Narens 2007), which strikes some to be too high a price to pay.

2.4 Relational species concepts

Another popular approach is to define a species as a group of individuals in a certain relationship to each other. The biological species concept, for instance, defines species as "groups of interbreeding populations that are reproductively isolated form other such groups (Mayr 1942)" so that the required relationship here is mutual crossability. Other variants focus on reproductive competition (Ghiselin 1974) or organisms' capacity to recognize each other as a possible mate (Paterson 1985). All these proposals try to reduce species into mutual relationships (interbreeding, competition, recognition, etc.) between a pair of organisms. If we represent such relationships by an edge between nodes/organisms, a relational species can be defined as an isolated complete subgraph or *clique* in an undirected graph, that is, a group of nodes in which every two distinct nodes are connected but none is connected to outside (Figure 1(d)). Relational species thus find their model in graph theory, where edges represent the relation in question.

A common criticism of relational species concepts is that the focal relationship such as crossability sometimes fails to induce isolated cliques because some organisms at a species boundary can often mate with organisms that are thought to belong another species (e.g. ring species). Moreover, the biological species concept has been criticized to imply every asexually reproducing organism forms a distinct species (for any singleton

node is complete). These criticisms suggest that the real biological network is so "messy" that just a single relationship cannot divide it into distinct cliques in a non-trivial way.

2.5 "Combo" solutions

The model-theoretic rendering makes explicit what each species concept can and cannot meaningfully say about the biological world. Given that most of the criticisms we have seen concern the "cannot say" part, one way to deal with these difficulties is to combine different theories to obtain more complex definitions of species.

For instance, one may combine the cluster and cladistic species concepts and define a species as a *lineage that shares the same or similar phenotypic distribution*:

$$\{\omega \in \Omega : \omega_f \prec \omega \wedge \theta(\omega_f) = \theta(\omega)\} \tag{2}$$

where $\theta : \Omega \to \mathbb{R}^n$ assigns distribution parameters to each organism $\omega \in \Omega$.[1] On this definition one may meaningfully define paraphyletic species and distinguish birds from other reptiles on the basis of the difference in their phenotypic or genetic profiles. It can also account for anagenesis (speciation without branching) and continuity of species between a cladogenesis (splitting event).

If one replaces θ in (2) with a different function $\nu : \Omega \to N$ that maps organisms $\omega \in \Omega$ to their *niche* $\nu(\omega) \in N$, it becomes the *ecological species concept* which defines a species as "a lineage ... which occupies an adaptive zone minimally different from that of

[1]For non-parametric cases, we can set $\theta : \Omega \to \mathbb{R}^\infty$ and modify the definition as $\{\omega \in \Omega : \omega_f \prec \omega \wedge D(\theta(\omega_f), \theta(\omega)) < k\}$ where $D(\bullet)$ is a divergence measure (such as the Kullback-Leibler divergence) and k is a constant.

any other lineage in its range (Van Valen 1976, 233)."

Yet another combination is that of the cladistic and biological species concepts, which would define a species as a maximum monophylic lineage that can mutually interbreed, so that

$$\{\omega_x, \omega_y \in \Omega : \omega_f \prec \omega_x \wedge \omega_f \prec \omega_y \wedge \omega_x \sim \omega_y\} \tag{3}$$

where $\sim$ stands for crossability.[2] This will make up for the lack of a sufficient condition in the cladistic species concept, and accord well with the so-called *evolutionary species concept* which emphasizes the unique "evolutionary tendencies and historical fate" of each species (Wiley 1978, 17). It should be noted that this could also avoid the problem of ring species because two crossable organisms may not necessary share the same ancestor.

2.6 The scientific species problem as a problem of theory choice

The above discussion shows that (i) major species concepts can be defined as models of formal theories, and that (ii) more complex concepts can be obtained by combining basic ones. The model-theoretic approach characterizes each species concept with the formal apparatus it assumes, which in turn determines its expressive power or what can meaningfully be stated about organisms and/or their history (Narens 2007). In general, a richer theoretical apparatus allows for more nuanced expressions, which makes it less liable to counterexamples. This is illustrated in the progression from the typological to

[2]As in the case of the biological species concept, the crossability here must take into account the existence of two sexes.

cluster and then to cluster-cladistic concepts, where in each step the species concept acquires the ability to deal with fuzzy boundaries and evolutionary history, respectively.

It does not necessarily follow, however, that a richer concept is always desirable, because it tends to have a greater degree of freedom and requires more data in actual application. While only phylogenetic information suffices to demarcate cladistic species, the cluster-cladistic concept also requires phenotypic or ecological information, which in many cases may not be available. A stronger semantic power thus comes with a higher epistemic cost, as is often emphasized by pheneticists or cladists in their respective advocacy of the phenotypic cluster and cladistic species concepts.

This suggests that the competition among various species concepts should be understood as a problem of model selection, where different models are evaluated on the basis of their explanatory or descriptive power versus parsimony or operationality (Sober 2008). Indeed, most disputes among advocates of different species concepts arise from their differential emphasis on what aspects of the biological world a disable species concept needs and needs not take into account (Ereshefsky 2001), but the difficulty is that these emphases are often implicit and incommensurable. Although the model-theoretic approach does not arbitrate these debates, it provides a common formal framework that makes explicit the explanatory power and operationality of species concepts and facilitates evaluation of their respective advantage.

3 Philosophical implications

3.1 Species are models

Upon the model-theoretic reconstruction of various species concepts, we now turn to the philosophical thesis that species taxa should be construed as models proposed above, i.e., as set-theoretic entities. To proceed, let me first begin with an analogy from classical mechanics. Classical mechanics is a theory about Newtonian particles, which are customary defined as volumeless points or vectors in a three-dimensional Euclidean space. Newton's celebrated laws like $\mathbf{F} = m\mathbf{a}$ describe temporal evolution of a system composed of such "particles." This system is to be distinguished from any actual physical systems, say the solar system, for one thing, no concrete bodies are volumeless, nor do they indefinitely continue rectilinear motion as prescribed by Newton's first law. Newton's theory, or any other physical theories for that matter, is a description of idealized and abstracted models and not of actual phenomena (Cartwright 1983). That is, models of classical mechanics — which make its laws and statements true — are not concrete, physical entities, but rather abstract mathematical objects that can be constructed within set theory (McKinsey et al. 1953).

The role of models in science has been emphasized by the so-called semantic or model-based view of scientific theories (e.g. van Fraassen 1980; Suppe 1989).[3] In the traditional, logical-positivist view, a scientific theory was supposed to directly describe

[3]This label ("the semantic view") has been used to describe different, and logically independent, theses. In particular, while some philosophers (e.g. Suppes 2002) take a scientific theory as a *description* of models, others *identify* it with a set of models (van Fraassen 1980). In this paper I adopt the former thesis without committing to the latter.

observed data. This has set for positivists the difficult task of reducing theoretical concepts that seemingly lack direct empirical contents to observation vocabulary by way of *bridge laws* or *partial interpretations*. To avoid this difficulty, proponents of the model-based view take a model, rather than observation, as the primary descriptive target of a scientific theory. In this view, a theory specifies an abstract model that idealizes and extracts just salient factors, and only indirectly relates to actual phenomena via such an model.

I submit that the species problem is a variant of the positivist conundrum. Species is a highly theoretical concept, and various proposal of "species concepts" in the past can be understood as attempts to build bridge laws for reducing it to a set of observational or operational criteria. To date more than a dozen of different concepts have been proposed[4], with no general consensus — each has its own strength, but also weakness and exceptions when applied to the rich and heterogeneous biological world. The assumption has been that a species concept must be a faithful description of *actual* biological features or phenomena. But what if this assumption is untenable, or at least unreasonable? The model-based view has been quite popular among philosophers of biology (e.g. Beatty 1981; Lloyd 1988). If we adopt this view and construe evolutionary theory as describing models, then species too must be defined accordingly, i.e., as (a part of) abstract models that satisfy descriptions and/or inferences of the corresponding theory.

What, then, are theories about species? Without claiming to be exhaustive, this paper adopts Suppes's (2002) thesis that a scientific theory must be defined as a set-theoretical predicate. The foremost advantage of this approach is that it enables one

[4]Mayden (1997), for example, counts at least 22 concepts of species.

to easily harness a theory with mathematical apparatus necessary for sophisticated reasoning. As discussed above, contemporary studies on species rely heavily on quantitative methods to calculate similarity or reconstruct a phylogenetic tree from phenotypic or genetic data. Given that such mathematical reasoning requires matching formal models of calculus or probability theory, the straightforward way to define a species is to build it upon these mathematical backgrounds as an extension of these formal models. Section 2 is a preliminary sketch of applying this Suppesian program to various species concepts. If this attempt turns out to be successful, biological species are to be understood as parts of set-theoretic structures, just like Newtonian particles. That is, they are mathematical and abstract constructs, rather than physical or biological entities.[5]

The purpose of the set-theoretic exposition is not just to accommodate quantitative reasoning. Even with less quantitative cases like the biological species concept, it makes implicit assumptions explicit and suggests a way to deal with counterexamples. The problem of ring species, for example, arises from a conflict between the presumption that each biological species must be isolated and the fact that crossability is not necessarily transitive and thus fails to induce equivalence classes. One possible response to this charge then would be to weaken the former assumption and redefine a species just as a (not necessarily isolated) clique in the reproductive network. Clarification of theoretical assumptions helps us to assess other species concepts as well. For example, the phenetic species concept is often claimed to be "theory-free" in that it does not depend on any evolutionary hypothesis. But as we have seen in Sec. 2.2, the calculation of phenotypic

[5]Hence the present thesis should not be confused with the view that species are sets or collections of *organisms* (Kitcher 1984), which, after all, are concrete biological entities.

similarity presupposes a phenotypic space equipped with a particular (e.g., euclidean) metric, which is a fairly strong theoretical assumption. Also, cladists often stress the simplicity and purity of their monophylic species definition that only considers phylogenetic relationships. But in order to make use of likelihood methods to infer such relationships, as is common in practice, a simple poset is not enough: one also needs to assume some genetic or phenotypic distribution, and then there is no in-principle reason to exclude non-monophylic taxa from the definition of species (as (2) in Sec. 2.5).

The final but not least merit of the set-theoretic approach is its flexibility: it allows for a construction of a new species concept by combining existing ones (Sec. 2.5) or adding new theoretical assumptions. For instance, it is common in experimental biology to characterize a species by shared developmental or causal mechanisms: developmental biologists often talk about "the development of the chicken" and medical doctors rely on causal extrapolation when they prescribe a clinically-tested drug for their patient. Such a "causal species" may be defined by isomorphic *causal models*, which combine a probabilistic distribution and a causal graph over variables. Hence the discussion in Section 2 covers just a few samples that can be constructed within this general framework. This does not of course mean that every possible species concept can and must be formalized, but does suggest the potential of the set-theoretic approach to accommodate the use of existing species concepts and to develop novel ones.

3.2 Philosophical implications

Identifying species with theoretical models sheds new light on some vexed philosophical issues, one amongst which concerns how individual organisms are related to species taxa.

Philosophers have long debated whether the organism-species relationship is instantial (organisms are particular *instances* of a species *qua* class), membership (they are *members* of a species *qua* set; Kitcher 1984), or mereological (they are *parts* of a species *qua* genealogical entity; Ghiselin 1997). The model-theoretic approach suggests an alternative account, according to which a species *represents* (a group of) individual organisms. Just as the Rutherford-Bohr model represents the microscopic structure of atoms, models proposed in Section 2 represent biological populations: for example, nodes and edges consisting of the biological species model in Figure 1(d) respectively represent organisms and crossability. Representation captures our intuitive notion that a model and its target phenomenon share salient static or dynamic features up to a certain precision. Given that said, it must be admitted that the criteria and nature of scientific representation are diversified and still open questions (Frigg and Nguyen 2016). Hence calling the species-organism relationship representational does not necessarily demystify it, but at least implies that the problem is not endemic to evolutionary theory: it is rather a version of a broader philosophical issue as to how the use of scientific models help us understanding the world. This means that the arsenal of this rich philosophical literature can and should be consulted to elucidate the nature of the species-organism relationship. Another, more immediate implication is that the membership and mereological accounts must be both abandoned, for whatever the relationship between a model and phenomena turns out to be, the latter must certainly not be a member or part of the former.

Neither is representation identity or instantiation. Ideal gas is not identical to any actual gas, but only approximates thermodynamic characteristics of some. Hence strictly speaking it has no instantiation, but this does not detract its epistemic validity. Likewise

species concepts, as specifications of ideal models, need not directly apply to actual populations. No wild population big enough to qualify as a species would strictly satisfy the requirement of the biological species concept, because actual mating chance is often hindered by physiological, geographical, and other contingencies. In the same vein, a phenetic or genetic cluster is expected to have outliers when applied to a real population. However, the presence of such exceptions should not immediately invalidate the corresponding species concepts, because the value of a species concept consists less in its universal validity than its epistemic serviceability for inferences and explanations of evolutionary or biological phenomena. These two criteria often conflict: Cartwright (1983) even argues that explanatory theories necessarily distort the reality by idealizing the situation and extracting only relevant features, so that properly speaking they are "lies" by design. Cartwright's examples are physics and economics, but her idea also applies to the present context. The primary function of a species concept is to explain biological phenomena rather than to save them, so that a few discrepancies should not be taken as a falsification.

The conflict between exceptionlessness versus explanatory power also underlies the realism-nominalism debate over species. The proponents of the nominalistic thesis who claim a species to be nothing but a totality of individual organisms have motivated their view by criticizing the realist interpretation of species-as-class for its commitment to the typological thinking and failure to deal with the evident heterogeneity of biological phenomena (e.g. Ghiselin 1997). On the other hand, those who attach weight on the role of species concept in induction and explanation have upheld a realist position and treated species as natural kinds (Boyd 1999). The present thesis offers a third alternative, recognizing the explanatory role of species concept without committing to

the ontologically heavy assumption of natural kinds. As we have seen in Section 2, species as models licence particular sets of inferences. The cluster and typological species/models underpin an expectation that physiological or genetic features found in, say, laboratory animals would also be shared by other individuals of the same species, while the evolutionary species concept explains the reason of such intra-specific similarities. These explanations are effectuated by the same model representing numerically distinct individuals or phenomena to be explained. Note that this procedure no more presupposes the existence of the model as an independent, real entity, than do explanations based on, say, ideal gas. Indeed, explanations may be based on fictional models, as is the case with the Ising model in statistical mechanics.

This does not of course mean that models *must be* fictions, or that species do not exist. Recent advocates of scientific realism argue that successful scientific models capture some, especially structural, aspect of reality (Ladyman 2016). Given its affinity to the model-based view of scientific theories, species realists may well apply this line of reasoning to the present context, taking the set-theoretic structures as discussed in Section 2 as representing the reality or "essential feature" of biological species. Whether and to what extent such an argument carry over, however, remain to be examined by a further study.

4 Conclusion

The past debates over biological species have been based on the assumption that species concepts must describe actual biological phenomena, the strict adherence to which tends to rule out all but cladistic species as typological or inexact. The present paper

challenged this assumption and argued that the primary referent of a species concept is a (set-theoretic) model that licences a certain set of inferences specified by the concept. The model-theoretic rendering articulates explanatory power and theoretical assumptions of each species concept and illuminates logical relationships among them. Once species are specified as models, the long-standing competition among different species concepts reduces to a common problem of model selection. This suggests that evaluation of relative merits and demerits of species concepts must be based more on their explanatory power than on exceptionlessness.

On the philosophical side, the shift in the ontological status of species means that the organism-species relationship is not that of instantiation, membership, or mereology, but rather representation. The vexed issue that has troubled philosophers for decades, therefore, boils down to the broader problem as to how and why scientific models can be used to represent and explain the world. This suggests the possibility to apply the rich literature on scientific representation and realism to elucidate the epistemological and ontological nature of biological species.

In sum, the take home message of the present paper is that the species problem is not endemic to biology or evolutionary theory, but rather is a variant of general scientific and philosophical issues of model selection, scientific representation, and realism. The purpose of this paper was just to establish such a parallelism: determining its philosophical implications on specific debates such as realism or pluralism concerning biological species will be a task for future studies.

References

Beatty, John. 1981. "What's Wrong with the Received View of Evolutionary Theory?." *PSA 1980* 2: 397–426.

Boyd, Richard N. 1999. "Homeostasis, species, and higher taxa." In *Species: New Interdisciplinary Essays.* ed. Robert A Wilson, 141–158, Cambridge, MA: MIT Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Ereshefsky, Marc. 2001. *The Poverty of the Linnaean Hierarchy.* Cambridge: Cambridge University Press.

van Fraassen, Bas C. 1980. *The Scientific Image.* Oxford: Oxford University Press.

Frigg, Roman, and James Nguyen. 2016. "Scientific Representation." In *The Stanford Encyclopedia of Philosophy.* ed. Edward N Zalta, Metaphysics Research Lab, Stanford University.

Ghiselin, Michael T. 1974. "A Radical Solution to the Species Problem." *Society of Systematic Biologists* 23: 536–544.

——— 1997. *Metaphysics and the Origin of Species.* Albany, NY: State University of New York Press.

Hennig, Willi. 1966. *Phylogenetic Systematics.* Urbana, IL: University of Illinois Press.

Hull, David L. 1976. "Are species really individuals?" *Systematic Zoology* 25: 174–191.

Kitcher, Philip. 1984. "Species." *Philosophy of Science* 51: 308–333.

Ladyman, James. 2016. "Structural Realism." In *The Stanford Encyclopedia of Philosophy.* ed. Edward N Zalta, Metaphysics Research Lab, Stanford University.

Lloyd, Elisabeth A. 1988. *The Structure and Confirmation of Evolutionary Theory.* Princeton, NJ: Princeton University Press.

Mayden, R L. 1997. "A hierarchy of species concepts: the denouement in the saga of the species problem." In *Species The Units of Biodiversity.* ed. M F Claridge, H A Dawah, and M R Wilson, 381–424, London: Chapman & Hall.

Mayr, Ernst. 1942. *Systematics and origin of species.* New York, NY: Columbia University Press.

McKinsey, John C C, Patrick Suppes, and A C Sugar. 1953. "Axiomatic Foundations of Classical Particle Mechanics." *Journal of Rational Mechanics and Analysis* 2: 253–272.

Narens, Louis. 2007. *Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science.* Mahwah, NJ: Lawrence Erlbaum Associates.

Paterson, Hugh E H. 1985. "The Recognition Concept of Species." In *Species and Speciation.* ed. E. S. Vrba, 21–29, Pretoria.

Sober, Elliott. 2008. *Evidence and Evolution.* Cambridge: Cambridge University Press.

Sokal, Robert R, and Peter H A Sneath. 1963. *Principles of Numerical Taxonomy.* San Francisco, CA: W. H. Freeman and Co.

Suppe, Frederick. 1989. *The Semantic Conception of Theories and Scientific Realism.*: University of Illinois Press.

Suppes, Patrick. 2002. *Representation and Invariance of Scientific Structures.* Stanford, CA: CSLI Publication.

Van Valen, Leigh. 1976. "Ecological Species, Multispecies, and Oaks." *Taxon* 25: 233–239.

Velasco, Joel D. 2008. "The internodal species concept: a response to 'The tree, the network, and the species'." *Biological Journal of Linnean Society* 93: 865–869.

Wiley, Edward O. 1978. "The Evolutionary Species Concept Reconsidered." *Systematic Biology* 27: 17–26.

Historical Inductions Meet the Material Theory

by Elay Shech

Oct. 2018
(Pre-conference version)

Forthcoming in *Philosophy of Science*

Acknowledgements: I am indebted to John Norton and Moti Mizrahi for extremely valuable discussion and comments on earlier drafts of this paper. Thank you also to helpful conversation with the audience at the Auburn University Philosophical Society in the Spring of 2018 and participants in Gila Sher's *Truth and Scientific Change* reading group in the Fall of 2017 at the Sidney M. Edelstein Center for History and Philosophy of Science, Technology and Medicine at the Hebrew University of Jerusalem.

Abstract: Historical inductions, viz., the pessimistic meta-induction and the problem of unconceived alternatives, are critically analyzed via John D. Norton's material theory of induction and subsequently rejected as non-cogent arguments. It is suggested that the material theory is amenable to a local version of the pessimistic meta-induction, e.g., in the context of some medical studies.

1. Introduction

My goal is to contribute to a growing literature that is critical of historical inductions such as the pessimistic (meta-)induction (PMI) argument (Poincaré 1952, 160; Putnam 1978, 25; Laudan 1981) and the problem of unconceived alternatives (Stanford 2001, 2006) against scientific realism, concentrating mostly on the former. The PMI can be construed in different ways (Mizrahi 2015, Wray 2015), viz., as a deductive *reductio ad absurdum* (e.g., Psillos 1996, 1999), a counterexample to the no miracles argument and inference to best explanation argument for scientific realism (e.g., Saatsi 2005, Laudan 1981), or, usually, as an inductive argument (e.g., Poincaré 1952, Putnam 1978, Laudan 1981, Rescher 1987). In the following I will argue against the inductive version of PMI—or any construal of the PMI that makes use of historical induction—using John D. Norton's material theory of induction (Norton 2003, Manuscript). The upshot is that one ought to be critical of historical inductions that seem to fit the general form or pattern of a good inductive argument, but may in fact lack inductive warrant and force. Various critiques have been put against the PMI (e.g., Lange 2002, Lewis 2001, Mizrahi 2013), along with some defenses (e.g., Saatsi 2005). In Section 2 I will present the PMI and briefly discuss some criticism in order to place my own analysis in broader context. Section 3 presents the material theory of induction and argues that it dissolves the PMI, while Section 4 extends such claims to the more recent problem of unconceived alternatives. In Section 5 I note that the material theory of induction does leave room for a local version of the PMI, which holds in some

limited domain, such as in relation to certain medical studies (Ruhmkorff 2014). I end in Section 6 with a short conclusion.

2. The (Inductive) Pessimistic (Meta-)Induction

The modern formulation of the PMI is usually attributed to Laudan (1981) who argued that having genuinely referential theoretical and observational terms, or being approximately true, is neither necessary nor sufficient for a theory being explanatory and predictively successful. More generally, Anjan Chakravartty characterizes the argument as follows:

> [PMI can] be described as a two-step worry. First, there is an assertion to the effect that the history of science contains an impressive graveyard of theories that were previously believed [to be true], but subsequently judged to be false . . . Second, there is an induction on the basis of this assertion, whose conclusion is that current theories are likely future occupants of the same graveyard. (Chakravartty 2008, 152)[1]

The PMI then may take the following form:

> **[Inductive Generalization PMI]**
>
> P(i) Past theory 1 was successful but not genuinely referential or approximately true.
> P(ii) Past theory 2 was successful but not genuinely referential or approximately true.
> …
> C) Therefore, current (and perhaps future) theories are successful but (by induction) probably not genuinely referential or approximately true.

Laudan (1981) suggests that the history of science contains a graveyard of theories that were previously believed to be approximately true and genuinely referential, but that subsequently were judged to be false and not to refer. Estimations of the number of such superseded theories have been debated (e.g., Lewis 2001, Wray 2013) and recently Mizrahi (2016) presents evidence that challenges the "history of science as a graveyard of theories" claim. Others voice concerns regarding the period of history of science used in order to extract historical evidence (e.g., Lange 2002, Fahrbach 2011) or the proper unit of analysis, i.e., theories vs. theoretical entity (e.g., Lange 2002, Magnus and Callender 2004). Similarly, Park (2011, 83) and Mizrahi (2013, 3220-3222) have argued that the PMI is fallacious due to cherry-picking data, biased statistics, and non-random sampling.

My own criticism of the inductive PMI comes from a different avenue. I will assume that the anti-realist does have randomly sampled historical evidence from the correct period of history and with the proper unit of analysis (whatever those

[1] cf. Wray (2015, 61).

may be) that is not biased or cherry-picked. Still, on the material theory of induction the PMI will not be a cogent argument. In other words, I aim to identify what I take to be a more fundamental (although not categorically different) problem with the PMI.

3. PMI Meets the Material Theory

3.1 The Material Theory of Induction in a Nutshell

Consider the following formally identical inductive inferences (Norton 2003, 649):

> P1) Some samples of the element bismuth melt at 271 degrees C.
> C1) Therefore, all samples of the element bismuth melt at 271 degrees C.
>
> P2) Some samples of wax melt at 91 degrees C.
> C2) Therefore, all samples of wax melt at 91 degrees C.

What makes the first argument an inductively strong and cogent argument while the second a weak and non-cogent inductive argument? Norton (2003, Manuscript) has argued that formal theories of induction, which provide universal schemas that are meant to identify the inductions that are licit and those that are not, stand against an insurmountable difficulty when facing such a question.[2] Instead, he offers a material account of induction:

> In a material theory, the admissibility of an induction is ultimately traced back to a matter of fact, not to a universal schema. We are licensed to infer from the melting point of some samples of an element to the melting point of all samples by a fact about elements: their samples are generally uniform in their physical properties. … *All inductions ultimately derive their licenses from facts pertinent to the matter of the induction*. (Norton 2003, 650; original emphasis)

Norton calls the local facts that power inductive inferences "material postulates." Material postulates themselves are supported by other instances of induction that are licensed by different material postulates.

3.2 Material Analysis of PMI

Many of the criticism of the inductive PMI discussed above amount to the claim that the universal schema used by the likes of Laudan (1981), namely, (P3) Some A's are B's, (C3) Therefore, all A's are B's, does not apply in the case of the PMI because various criteria needed to implement the scheme, e.g., random sampling, correct historical period, proper unit of analysis, have not been met. What I wish to do here

[2] I will not defend Norton's theory or claims here. He dedicates an entire book to the matter in Norton (Manuscript).

is conduct a material analysis of the PMI. Considering the above presentation of the PMI in its [Inductive Generalization PMI] form we may ask, what powers the inductive inference, i.e., what material postulate licenses the pessimistic conclusion?

In context of the two inductive arguments considered in Section 3.1, we note that there is no material postulate that licenses the inductive inference in the case of wax (P2 too C2) but there is one in the case of bismuth (P1 to C1): Generally, chemical elements are uniform in their physical properties. By analogy, the presumption of the meta-induction is that each historical case study looked at is an instance of the same thing, a discovery of induction in science. If we are to perform the meta-induction then there needs to be something in the background facts that unifies all such inductions, just like the fact chemical elements are generally uniform in their physical properties warrants the inductive inference regarding the melting point of bismuth. Let us consider several options.

First, perhaps the material fact is that most scientists use a common rule or method in constructing or discovering successful theories, something along the lines of Mill's methods of experimental inquiry in his *System of Logic* (1872, Book III, Ch. 7). If so, the properties of the rule would be used to authorize the induction. Is there such a rule, or perhaps, some common scientific method? A glance at the history of science suggests that this is unlikely. Newton's deduction from the phenomena, is very different from Darwin's inference to best explanation, which in turn differs radically from Einstein's thought experiments with lights beams, trains, and elevators.[3] More generally, there seems to be a consensus among historians and philosophers of science that something like "the scientific method" is really more of an umbrella term for very different methods used by scientists to construct and discover theories. After all, novel problems necessitate novels solutions, and the commonality that does arise in different cases, say, attempts to minimize error or to be objective, is not the kind of commonality that we seek in powering the PMI and drawing the pessimistic conclusion. For instance, in his book *Styles of Knowing: A New History of Science from Ancient Times to the Present*, Chungling Kwa (2011) argues that there is no single, fundamental method used in science: "there is not just one form of Western scientific rationality; there are at least six." The framework of six "styles of knowing," includes the deductive, the experimental, the hypothetical-analogical, the taxonomic, the statistical, and the evolutionary style, and is based on Alistair Crombie's (1994) three-volume work *Styles of Scientific Thinking*. Similar, Ian Hacking (also taking lead from Crombie's work) has argued that there are distinct "styles of reasoning" used in science, such as the postulational style, the style of experimental exploration, the style of hypothetical construction of models by analogy, the taxonomic style, the statistical style, the historical derivation of genetic development, and the laboratory style (Hacking 1992). This further

[3] In fact, see Norton (Manuscript, Ch. 8-9) who argues that even in historical cases where the *same* principle is applied by scientists, viz., inference to best explanation, "at best we can find loose similarities that the canonical examples of inference to best explanation share," so that no common rule of the kind needed to power the PMI can be found (Ch. 8, p. 1).

corroborates the idea that scientific methods used for theory construction and discovery, as well as for scientific explanation, are very diverse.

More generally, scientific theories are not kind of things that portray the type of uniformity needed to license inductive inferences on Norton's material theory. Albeit in a different context, a similar point is nicely made by Mizrahi (2013, 3218):

> A uniform—as opposed to diverse—sample might be a sample of, say, copper rods. From a sample of just a few copper rods that are tested for electrical conductivity, it is reasonable to conclude that all copper rods conduct electricity because, if you have seen one or two copper rods, you have seen them all (given their uniform atomic structure). Scientific theories, however, are not as uniform as copper rods. The point, then, is that any sample of theories is not going to be uniform in a way that is required for a "seen one, seen them all" inductive generalization.

Similarly, and second, perhaps there are some facts about investigating scientist themselves, how they work, and/or the problems situations that they work in, which can unify the historical evidence in a way that provides us with the inductive warrant we seek. Maybe such facts will include something about the psychology of scientists: their fastidiousness and fear of error, their facility at jumping to conclusions, or perhaps their curiosity, logic, creativity, skepticism, etc. However, in a similar manner to the search for a common rule used in constructing successful theories, the history of science furnishes us with scientists that are heterogeneous enough in their psychological traits, and work in such varied contexts, so as not to provide us with any was to unify the various historical cases in a way pertinent to licensing the pessimistic inference of the PMI.

Third, perhaps we can circumvent looking to a common rule of constructing or discovering theories, or searching for common traits among scientists, by noting that the follow candidate material postulate would power the PMI:

> MP-PMI: Generally, successful theories are not genuinely referential and/or approximately true.

But how would we establish MP-PMI? One option is to appeal to the PMI itself, but this would either be circular or else push us to look for another material postulate. Another option is just to grant the MP-PMI as a reasonable assumption. Perhaps anti-realists or instrumentalists would think that this is a sensible starting point, but their target realist opponent would surely reject such an assumption as question begging. Last, perchance there is some fact about explanatory and/or predictively successful theories that renders them, generally, not genuinely referential and/or approximately true? Possibly part of the essence of successful theories is to misrepresent the world? To me this seems highly unlikely and at odds with any levelheaded intuition but, in any case, if we could argue that successful theories are essentially inaccurate then we would not need the PMI in the first place!

Fourth, we may want to construe the PMI in its inductive generalization form as a kind of abductive argument with the following type of material postulate:[4]

> **[Inductive Generalization PMI – Abductive version]**
>
> P(i): The success of past theory 1 (constructed using method m) is not best explained by its truth.
> P(ii): The success of past theory 2 (constructed using method m) is not best explained by its truth.
> …
> MP: Scientific theories constructed using method m are generally uniform with respect to what best explains their predictive success.
> C: The success of our best current (and perhaps futures) theories (constructed using method m) are not best explained by their truth.

Stating the PMI as above has the merit of directly engaging with the "no miracles argument" for scientific realism, namely:

> That terms in mature scientific theories typically refer [to things in the world] …, that theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when it occurs in different theories—these statements are viewed by the scientific realist not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate scientific description of science and its relations to its objects. (Putnam 1975, 73)

But worries abound. First, the realist may very well deny P(i), P(ii), etc., and argue that the success of past theories is best explained by their truth but that, as it turns out, either the best explanation did not hold in this case or else there is some sense in which past theories, insofar as they were successful, were approximately true or on the road to truth. Second, construing the argument as an abduction opens up a Pandora's box of problems associated with the notion of explanation: What is explanation? Are there accounts of explanation where success is best explained by truth and ones in which it isn't and, if so, which account of explanation is relevant in this context? And so on.

Third, the cogency of the argument depends on the idea that all theories appealed to were constructed with some method m, but we already judged that there is no one method that is relevant to constructing scientific theories. Perhaps phenomenological models are good candidates for the type of things that can provide empirical success but are not generally approximately true.[5] Thus, at best, the above argument can power a kind of local PMI: Successful theories constructed

[4] Thanks to Tim Sundell for suggest this line of thought.
[5] Phenomenological models are, generally, not considered explanatory.

by method m are not approximately true. We'll consider one such case in more detail in Section 5.

In short, on the material theory of induction inductive arguments are powered by facts, by material postulates, but in the context of the PMI it seems unlikely that any such non-question begging postulates, which wouldn't render the PMI obsolete, can be found. This is so even if, say, the historical data was not cherry-picked, and the right unit of analysis and correct period of history were used. In other words, I'm equally skeptic of projects that attempt to block the pessimistic conclusion by, for example, taking a random sample of past scientific theories, e.g., Mizrahi (2016). In the following section I'll attempt to extend such claims to the problem of unconceived alternatives.

4. Extension to the Problem of Unconceived Alternatives

Recently, P. Kyle Stanford (2001, 2006) has developed what may be characterized as a new version of the PMI:

> … I propose the following New Induction over the History of Science: that we have, throughout the history of scientific inquiry and in virtually every field, repeatedly occupied an epistemic position in which we could conceive of only one or a few theories that were well-confirmed by the available evidence, while subsequent history of inquiry has routinely (if not invariably) revealed further, radically distinct alternatives as well-confirmed by the previously available evidence as those we were inclined to accept on the strength of that evidence. (Stanford 2001, S8-S9)

The problem of unconceinved alternatives as an argument against scientific realism has been criticized on various grounds (e.g., Chakravartty 2008, Devitt 2011, Mizrahi 2015), but my goal here is just to note that the discussion of Section 3 can be extended to this new version of the PMI, which can be construed as follows:

> P(i) In the past time of theory 1, theory 1 was successful but there were unconceived alternative theories that were as well supported by available evidence but with radically different ontology.
> P(ii) In the past time of theory 2, theory 2 was successful but there were unconceived alternative theories that were as well supported by available evidence but with radically different ontology.
> …
> C) Therefore, in present times, current theories are successful but (by induction) there probably are unconceived alternative theories that are as well supported by available evidence but with radically different ontology.

What we need for the material analysis is something like: Generally, successful theories are underdetermined by data due to possible unconceived alternative theories. In a similar fashion to the MP-PMI, we could look to some common rule used by scientists to conceive theories, or some common psychological traits among

scientist, that may ground the idea that successful theories are such that empirically adequate unconceived alternatives always exists. But for the same reasons discussed above, it seems unlikely that any such common rule or traits will be found. That said, perhaps cognitive facts about human scientists might support the inductive inference to the conclusion that we always miss some alternative theories, which in turn are consistent with the available evidence. What is attractive about this line of thought is that it does seem plausible that due to our cognitive limitations there are always "unconceived alternatives." However, mere cognitive limitations do not support the further conclusion that there are unconcieved alternative theories that are *consistent with available evidence*.

Alternatively, one may think that Stanford's new induction circumvents the material objection: modal reflections alone convince us that there are always unconceived alternative theories that can explain and predict empirical phenomena just as well or better than conceived theories. But how can we come to such a conclusion based on modal reflections alone? Isn't it conceivable if not possible that there would be a point in history with no unconceived alternatives and isn't conceivable if not possible that we are at such point in time in history? Moreover, it is unclear what to make of theory-independent modal claims (unless one has logical modality in mind, which isn't the case here). Certainly, we can talk about different physically possible worlds given a particular physical theory. For instance, various solutions to the Einstein field equations are taken to denote different possible universes according to relativity theory. But it isn't clear what is meant by different possible or alternative conceivable *theories* given no meta-theory as a constraint, so to speak.[6] In any case, if we know that unconceived alternative theories always exist based on modal reflections alone, then the historical induction is doing no work for us at all.

5. Room for a local, material pessimistic induction?

Although the material analysis given here may prompt us to be skeptical of historical inductions (insofar as one is moved by the material theory of induction), it can help us understand why *local* pessimistic inductions may be tenable. Specifically, I want to look at a recent discussion by Rumkorf (2014) who contends that meta-analyses in medicine such as Ioannidis' (2005a, 2005b), which show that a disconcertingly high percentage of prominent medical research findings are refuted by subsequent research, can be developed into a local pessimistic induction. Ioannidis (2005a, 2005b) is concerned with studies, denoted "M-studies," that satisfy the following criteria: "being highly cited, using contemporary research and statistical methods, and being among the first studies to investigate a question at issue" (Rumkorf 2014, 420). Rumkorf's (2014, 421) then uses the various conclusions of Ioannidis (2005a, 2005b) to generate a local PMI in the field of medicine (PMI-M):

[6] What would count as a (logically possible but physically) impossible theory in such a context?

> E1 41% of the associative or causal claims made by M-studies in the sample were inconsistent with the results of subsequent published studies either (1) because the later studies provided evidence against the existence of the association or effect; or (2) because the later studies provided evidence that the magnitude of the association or effect was significantly different.
>
> E2 Therefore, we can expect approximately 41% of the associative and causal claims made by M-studies to be inconsistent with subsequent published studies.

On Norton's theory we need to appeal to a material postulate to license the pessimistic inductive inference in the transitions from E1 to E2, but since we are now working in a limited domain without many heterogeneous examples as in the whole history of science, we may now find some significant commonality between the methods used in different M-studies that can act as licensing facts. What are the background facts that power the PMI-M? Here are some options extracted from Ioannidis's diagnosis of his meta-analysis and quoted in Ruhmkorf (2014, 219):

> Contributing factors include: bias in research (Ioannidis 2005b); non-randomized trials (Ioannidis 2005a); smaller rather than larger sample sizes in refuted studies (Ioannidis 2005a, 224); and publication and time-lag biases (whereby studies with highly significant and potentially aberrational positive results are overrepresented among published articles in major journals and are published more quickly than other articles) (Ioannidis 2005a, 224). Particularly intriguing is the idea that large-scale features of the structure of medical and biological inquiry contribute to the high contradiction rate. Having a number of distinct working groups looking at the same problem increases the chances that at least one of them will find something statistically significant, especially if they are looking at a wide array of possible relationships (Ioannidis 2005b, 697–698). The computational power and richness of data sets available to researchers increases the chance that some of them will be successful in achieving statistical significance, even when no real relationship exists (Ioannidis 2005b, 701).[7]

These various factors, insofar as they are common to most M-studies, are the type of background facts that warrant the pessimistic induction from a material point of view. One may worry of course that the pessimism associated with local PMI generalizes since, presumably, facts about biases and the like are facts about researchers in general, not just researchers in medical science in particular. But, although all scientific studies have to deal with challenges such bias, it may be the case that a particular local subfield, due to its specific nature and whatever social

[7] It should be noted that there are some problems with Ioannidis's (2005a, 2005b) methodology, as identified in Ruhmkorff (2014, 419-421), but they do not seem to be problematic enough to render the PMI-M not cogent.

norms are in place for collecting and disseminative evidence, is especially challenged in a way that can justify the pessimistic induction. The above suggests that this is indeed the case for M-studies.

To end, Ruhmkorff (2014) argues against global PMI on independent grounds (namely, he argues that the PMI commits a statistical error previously unmentioned in the literature and is self-undermining), and but he also argues for the plausibility of a local PMI, viz., M-PMI, and contends that there are clear advantages of PMI-M over PMI. What I wish to note here is that an additional advantage of PMI-M, or local pessimistic induction generally speaking, is that whereas global PMI dissolves upon a material analysis, a material account of PMI-M does seem viable.

6. Conclusion

I have argued that historical inductions such as the (global) PMI and the problem of unconceived alternatives dissolve if we work with the material theory of induction. The reason is that we lack the material postulates needed to license the pessimistic inference: the great heterogeneity of case studies from the history of science of conceiving, constructing, and discovering (explanatory and predictively successful) theories, along with abundant variety of context that scientists find themselves in and traits that they exhibit, make it unlikely that any commonality will be found strong enough to authorize the induction. One may of course object: so much worse for the material theory of induction! This is a fair point, but there is a more general moral to consider. In various situations one may be able to appeal to the notion of "induction" without much being at stake, but in the context of historical inductions like the PMI and problem of unconceived alternatives "induction" is doing a lot of (philosophically) heavy lifting and so the situation rightful calls for scrutiny. Such scrutiny has led to the various discussed criticism that are presented in the context of more traditional, non-material theories of induction. Accordingly, it seems appropriate to show that—even if we assume randomly sampled historical evidence from the correct period of history and with the proper unit of analysis that is not biased or cherry-picked, with no statistical error, etc.—historical inductions do not fare well on the material side of things. I leave objections to the effect that one ought to construe the PMI as a deductive argument, or through a different framework for induction, e.g., via hypothetical or probabilistic induction, for future work.

References

Chakravartty, A. 2008. "What You Don't Know Can't Hurt You: Realism and the Unconceived." *Philosophical Studies* 137: 149–158.

Crombie, A. C., 1995. *Styles of Scientific Thinking in the European Tradition*, 3 vols. London: Duckworth.

Devitt, M. 2011. "Are Unconceived Alternatives a Problem for Scientific Realism?" *Journal for General Philosophy of Science* 42: 285–293.

Fahrbach, L. 2011. "How the Growth of Science Ends Theory Change." *Synthese* 180: 139–155.

Hacking, I. 1992. "'Style' for historians and philosophers." *Studies in History and Philosophy of Science*, 23(1), 1–20.

Ioannidis, J. P. A. 2005a. "Contradicted and Clinically Stronger Effects in Highly Cited Clinical Research." *Journal of the American Medical Association* 294: 218–228.

Ioannidis, J. P. A. 2005b. "Why Most Published Research Findings Are False."*PLoS Medicine 2*: 696–701.

Kwa, C. 2011. *Styles of Knowing: A New History of Science from Ancient Times to the Present*. Pittsburgh: University of Pittsburgh Press.

Lange, M. 2002. "Baseball, Pessimistic Inductions, and the Turnover Fallacy." *Analysis* 62: 281–285.

Laudan, L. 1981. "A Confutation of Convergent Realism." *Philosophy of Science* 48: 19–49.

Lewis, P. J. 2001. "Why the Pessimistic Induction Is a Fallacy." *Synthese* 129: 371–380.

Magnus, P. D., and C. Callender. 2004. "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71: 320–338.

Mill, J. S. [1872] 1916. *A System of Logic: Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. 8th ed. London: Longman, Green, and Co.

Mizrahi, M. 2013. "The Pessimistic Induction: A Bad Argument Gone too Far." *Synthese* 190:3209–3226.

Mizrahi, M. 2015. "Historical Inductions: New Cherries, Same Old Cherry-picking." *International Studies in the Philosophy of Science* 29: 129–148.

Mizrahi, M. 2016. "The history of Science as a Graveyard of Theories: A Philosophers' Myth?" *International Studies in the Philosophy of Science* 30: 263-278.

Norton, J. D. 2003. "A Material Theory of Induction." Philosophy of Science 70: 647–670.

Norton, J. D. Manuscript. *The Material Theory of Induction*. See http://www.pitt.edu/~jdnorton/papers/material_theory/material.html

Park, S. 2011. "A Confutation of the Pessimistic Induction." *Journal for General Philosophy of Science* 42: 75–84.

Poincaré, H. [1902] 1952. *Science and Hypothesis*. New York: Dover. Originally published as *La science et l'hypothèse*. Paris: Flammarion.

Putnam, H. 1978. Meaning and the Moral Sciences. London: Routledge and Kegan Paul.

Psillos, S.: 1996, 'Scientific Realism and the 'Pessimistic Induction' ', *Philosophy of Science* 63 (Proceedings), S306–S314.

Psillos, S. 1999. Scientific Realism: How Science Tracks Truth. London: Routledge.

Rescher, N. 1987. Scientific Realism: A Critical Reappraisal. Dordrecht: D. Reidel.

Ruhmkorff, S. 2013. "Global and Local Pessimistic Meta-inductions." *International Studies in the Philosophy of Science* 27: 409–428.

Saatsi, J. 2005. "On the Pessimistic Induction and Two Fallacies." *Philosophy of Science* 72: 1088–1098.

Sklar, L. M. (2003). "Dappled theories in a uniform world." *Philosophy of Science*, 70, 424–441.

Stanford, P. K. 2001. "Refusing the Devil's Bargain: What Kind of Underdetermination Should We take Seriously?" *Philosophy of Science* 68 (Proceedings): S1–S12.

Stanford, P. K. 2006. Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford: Oxford University Press.

Wray, K. Brad. 2015. "Pessimistic Inductions: Four Varieties." *International Studies in the Philosophy of Science* 29: 61–73.

PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association -510-

To be presented at the *2018 PSA Meeting*:

Can Quantum Thermodynamics Save Time?

Noel Swanson*

Abstract

The *thermal time hypothesis (TTH)* is a proposed solution to the problem of time: every statistical state determines a thermal dynamics according to which it is in equilibrium, and this dynamics is identified as the flow of physical time in generally covariant quantum theories. This paper raises a series of objections to the TTH as developed by Connes and Rovelli (1994). Two technical challenges concern the implementation of the TTH in the classical limit and the relationship between thermal time and proper time. Two more conceptual problems focus on interpreting the flow of time in non-equilibrium states and the lack of gauge invariance.

1 Introduction

In both classical and quantum theories defined on fixed background spacetimes, the physical flow of time is represented in much the same way. Time translations correspond to a continuous 1-parameter subgroup of spacetime symmetries, and the dynamics are implemented either as a parametrized flow on statespace (Schödinger picture) or a parametrized group of automorphisms of the algebra of observables (Heisenberg picture). In generally

*Department of Philosophy, University of Delaware, 24 Kent Way, Newark, DE 19716, USA, nswanson@udel.edu

covariant theories, where diffeomorphisms of the underlying spacetime manifold are treated as gauge symmetries, this picture breaks down. There is no longer a canonical time-translation subgroup at the global level, nor is there a gauge-invariant way to represent dynamics locally in terms of the Schrödinger or Heisenberg pictures. Without a preferred flow on the space of states representing time, the standard way to represent physical change via functions on this space taking on different values at different times, also fails. This is the infamous *problem of time*.

Connes and Rovelli (1994) propose a radical solution to the problem: the flow of time (not just its direction) has a thermodynamic origin. Equilibrium states are usually defined with respect to a background time flow (e.g., dynamical stability and passivity constraints reference a group of time translations). Conversely, given an equilibrium state one can derive the dynamics according to which it is in equilibrium. Rovelli (2011) exploits this converse connection, arguing that in a generally covariant theory, *any* statistical state defines a notion of time according to which it is an equilibrium state. The *thermal time hypothesis (TTH)* identifies this state-dependent thermal time with physical time. Drawing upon tools from Tomita-Takesaki modular theory, Connes and Rovelli demonstrate how the TTH can be rigorously implemented in generally covariant quantum theories.

The idea is an intriguing one that, to date, has received little attention from philosophers.[1] This paper represents a modest initial attempt to sally forth into rich philosophical territory. Its goal is to voice a number of technical and conceptual problems faced by the TTH and to highlight some tools that the view has at its disposal to respond.

2 The Thermal Time Hypothesis

We usually think of theories of mechanics as describing the evolution of states and observables through time. Rovelli (2011) advocates replacing this picture with a more general *timeless* one that conceives of mechanics as describing relative correlations between physical quantities divided into two classes, *partial* and *full* observables. Partial observables are quantities that physical measuring devices can be responsive to, but whose value cannot be predicted

[1] Earman (2002), Earman (2011), and Ruetsche (2014) are notable exceptions. Physicists have been more willing to dive in. Paetz (2010) gives an excellent critical discussion of the many technical challenges faced by the TTH.

given the state alone (e.g., proper time along a worldline). A full observable is understood as a coincidence or correlation of partial observables whose value can be predicted given the state (e.g., proper time along a worldline at the point where it intersects another worldline). Only measurements of full observables can be directly compared to the predictions made by the mechanical theory.

A timeless mechanical system is given by a triple $(\mathcal{C}, \Gamma, f)$. $\mathcal{C}$ is the configuration space of partial observables, q^a. A *motion* of the system is given by an unparametrized curve in $\mathcal{C}$, representing a sequence of correlations between partial observables. The space of motions, Γ is the statespace of the system and is typically presymplectic. The evolution equation is given by $f = 0$, where f is a map $f : \Gamma \times \mathcal{C} \to V$, and V is a vector space. For systems that can be modeled using Hamiltonian mechanics, Γ and f are completely determined by a surface Σ in the cotangent bundle $T^*\mathcal{C}$ (the space of partial observables and their conjugate momenta p_a). This surface is defined by the vanishing of some Hamiltonian function $H : T^*\mathcal{C} \to \mathbb{R}$.

If the system has a preferred external time variable, the Hamiltonian can be decomposed as

$$H = p_t + H_0(q^i, p_i, t) \tag{1}$$

where t is the partial observables in $\mathcal{C}$ that corresponds to time. Generally covariant mechanical systems lack such a canonical decomposition. Although these systems are fundamentally timeless, it is possible for a notion of time to emerge thermodynamically. A closed system left to thermalize will eventually settle into a time-independent equilibrium state. Viewed as part of a definition of equilibrium, this thermalization principle requires an antecedent notion of time. The TTH inverts this definition and use the notion of an equilibrium state to select a partial observable in $\mathcal{C}$ as time.

Three hurdles present themselves. The first is providing a coherent mathematical characterization of equilibrium states. The second is finding a method for extracting information about the associated time flow from a specification of the state. Finally, in order to count as an emergent explanation of time, one has to show that the partial observable selected behaves as a traditional time variable in relevant limits.

For generally covariant quantum theories, Connes and Rovelli (1994) propose a concrete strategy to overcome these hurdles. Minimally, such a theory can be thought as a non-commutative C^*-algebra of diffeomorphism-invariant

observables, $\mathfrak{A}$, along with a set of physically possible states, $\{\phi\}$.[2] Via the Gelfand-Nemark-Segal (GNS) construction, each state determines a concrete Hilbert space representation $(\pi_\phi(\mathfrak{A}), \mathcal{H}_\phi)$, and a corresponding von Neumann algebra $\pi_\phi(\mathfrak{A})''$, defined as the double commutant of $\pi_\phi(\mathfrak{A})$.

Connes and Rovelli first appeal to the well-known *Kubo-Martin-Schwinger (KMS) condition* to characterize equilibrium states. A state, ρ, on a von Neumann algebra, $\mathfrak{M}$, satisfies the KMS condition for inverse temperature $0 < \beta < \infty$ with respect to a 1-parameter group of automorphisms, $\{\alpha_t\}$, if for any $A, B \in \mathfrak{M}$ there exists a complex function $F_{A,B}(z)$, analytic on the strip $\{z \in \mathbb{C} | 0 < \mathrm{Im} z < \beta\}$ and continuous on the boundary of the strip, such that

$$\begin{aligned} F_{A,B}(t) &= \rho(\alpha_t(A)B) \\ F_{A,B}(t + i\beta) &= \rho(B\alpha_t(A)) \end{aligned} \tag{2}$$

for all $t \in \mathbb{R}$. The KMS condition generalizes the idea of an equliibirum state to quantum systems with infinitely many degrees of freedom. KMS states are stable, passive, and invariant under the dynamics, $\{\alpha_t\}$. Moreover in the finite limit, the KMS condition reduces to the standard Gibbs postulate.

Although the KMS condition is framed relative to a chosen background dynamics, according to the main theorem of *Tomita-Takesaki modular theory*, every faithful state determines a canonical 1-parameter group of automorphisms according to which it is a KMS state. Connes and Rovelli go on to identify the flow of time with the flow of this state-dependent *modular automorphism group*.

In the GNS representation $(\pi_\phi(\mathfrak{A}), \mathcal{H}_\phi)$, the defining state, ϕ, is represented by a cyclic vector $\Phi \in \mathcal{H}_\phi$. If ϕ is a *faithful* state (i.e., if $\phi(A^*A) = 0$ entails that $A = 0$) then the vector Φ is also separating. In this setting we can apply the tools of Tomita-Takesaki modular theory. The main theorem asserts the existence of two unique modular invariants, an antiunitary operator, J, and a positive operator, Δ. (Here we will only be concerned with the latter.) The 1-parameter family, $\{\Delta^{is} | s \in \mathbb{R}\}$, forms a strongly continuous unitary group,

$$\sigma_s(A) := \Delta^{is} A \Delta^{-is} \tag{3}$$

for all $A \in \pi(\mathfrak{A})''$, $s \in \mathbb{R}$. The defining state is invariant under the flow of the modular automorphism group, $\phi(\sigma_s(A)) = \phi(A)$. Furthermore, $\phi(\sigma_s(A)B) =$

[2] See Brunetti et al. (2003) for a formal development of this basic idea.

$\phi(B\sigma_{s-i}(A))$. Thus ϕ satisfies the KMS condition relative to $\{\sigma_s\}$ for inverse temperature $\beta = 1$.

For any faithful state, this procedure identifies a partial observable, the thermal time, $t_\phi := s$, parametrizing the flow of the (unbounded) thermal hamiltonian $H_\phi := -\ln \Delta$, which has Φ as an eigenvector with eigenvalue zero. We can then go on to decompose the timeless Hamiltonian $H = p_{t_\phi} + H_\phi$. Associated with any such state, there is a natural "flow of time" according to which the system is in equilibrium. But in what sense does this thermal time flow correspond to various notions of physical time? In particular, how is thermal time related to the proper time measured by a localized observer?

Although they do not establish a general theorem linking thermal time to proper time, Connes and Rovelli do make substantial progress on the third hurdle in one intriguing special case. For a uniformly accelerating, immortal observer in Minkowski spacetime, the region causally connected to her worldline is the *Rindler wedge*. In standard coordinates we can explicitly write the observer's trajectory as

$$\begin{aligned} x^0(\tau) &= a^{-1}\sinh(\tau) \\ x^1(\tau) &= a^{-1}\cosh(\tau) \\ x^2(\tau) &= x^3(\tau) = 0 \end{aligned} \tag{4}$$

where τ is the observer's proper time. The wedge region is defined by the condition $x^1 > |x^0|$. The *Bisognano-Wichmann theorem* then tells us that in the vacuum state, the modular automorphism group for the wedge implements wedge-preserving Lorentz boosts — Δ^{is} is given by the boost $U(s) = e^{2\pi i s K_1}$ (where K_1 is the representation of the generator of an x^1-boost). Since the Lorentz boost $\lambda(a\tau)$ implements a proper time translation along the orbit of an observer with acceleration a, $U(\tau) = e^{ai\tau K_1}$ can be viewed as generating evolution in proper time. Comparing these two operators, we find that proper time is directly proportional to thermal time,

$$s = \frac{2\pi}{a}\tau \tag{5}$$

The Unruh temperature measured by the observer is $T = a/2\pi k_b$ (where k_b is Boltzmann's constant), this leads Connes and Rovelli to propose that the Unruh temperature can be interpreted as the ratio between thermal and proper time. Not only does this relationship hold along the orbits of constant

acceleration, but if an observer constructs global time coordinates for the wedge via the process of Einstein synchronization, this global time continues to coincide with the rescaled thermal time flow.

We can now summarize the main content of the TTH:

Thermal Time Hypothesis (Rovelli-Connes)**.** *In a generally covariant quantum theory, the flow of time is defined by the state-dependent modular automorphism group. The Unruh temperature measured by an accelerating observer represents the ratio between this time and her proper time.*

This is a bold idea with a numerous potential implications for quantum physics and cosmology. Over the next three sections, we will consider a series of technical and conceptual objections to the TTH.

3 Thermal Time and Proper Time

The Bisognano Wichmann theorem only applies to immortal, uniformly accelerating observers in the vacuum state of a quantum field theory in flat spacetime. How can we characterize the relationship between thermal and proper time for a broader, more physically realistic class of observers and theories?

A uniformly accelerating mortal observer has causal access to a different region of Minkowski spacetime, the *doublecone* formed by the intersection of her future lightcone at birth and her past lightcone at death. Because wedges and doublecones can be related by a conformal transformation, in conformally invariant theories, geometric results from wedge algebras can be transferred onto the doubelcone algebras. In the vacuum state of a conformal theory, the doublecone modular automorphism group acts as Hislop-Longo transformations (Hislop and Longo, 1982). Martinetti and Rovelli (2003) use this result to calculate the corresponding relationship between thermal time and proper time for a uniformly accelerating mortal observer:

$$s = \frac{2\pi}{La^2}(\sqrt{1 + a^2L^2} - \cosh a\tau) \tag{6}$$

where L is the observer's lifetime. (The relationship is more complicated in this case due to the fact that proper time is bounded while modular time is unbounded.) For most of the observer's lifespan, s is an approximately constant function of τ, allowing the Unruh temperature to again be interpreted as the local ratio between thermal and proper time.

This is the best we can hope for. Trebels (1997) proves that arbitrary doublecone automorphisms act as local dynamics, only if they act as scaled Hislop-Longo transformations.[3] Of course, if nature is described by a non-conformal theory, then there is no guarantee that the doublecone modular automorphisms will have a suitable geometric interpretation. Saffary (2005) goes further, arguing that they will not have geometric significance in any theory with massive particles. The mathematical results backing this conjecture, however, are only partial.[4]

Attempting to generalize the TTH to cover non-uniform acceleration and non-vacuum states generates further difficulties. Work on the Unruh effect for non-uniformly accelerating observers (e.g., Jian-yang et al. 1995), indicates that such observers feel an acceleration-dependent thermal bath, reflecting the shifting ratio between constant thermal time and acceleration-dependent proper time. The TTH must explain the phenomenological experience of the observer who will presumably age according to her proper time, not the background thermal time flow. On top of this, if the global state is not a vacuum state, then it is not clear that the wedge modular automorphisms will carry a dynamical interpretation at all. The Radon-Nikodym theorem ensures that the action of the modular automorphism group uniquely determines the generating state. If ϕ, ψ are two (faithful, normal) states on a von Neumann algebra $\mathfrak{M}$, then the associated modular automorphism groups σ^t_ϕ, σ^t_ψ differ by a non-trivial inner automorphism, $\sigma^t_\phi(A) = U\sigma^t_\psi(A)U^*$, for all $A \in \mathfrak{M}$, $t \in \mathbb{R}$, so the general wedge dynamics will not be simple rescalings of the vacuum case.

None of these are knockdown objections since so little is known about the geometric action of modular operators apart from the Bisognano-Wichmann theorem and its conformal generalization. But our current ignorance also presents a major challenge. (The situation is even less clear in general curved spacetime settings.) The defender of the TTH has at least four options on

[3]Formally, Trebels requires that local dynamics be continuous 1-parameter groups of automorphisms of the doublecone algebra that preserve subalgebra localization as well as spacelike and timelike relations between interior points. For a detailed discussion of Trebels's results, see Borchers (2000), §3.4.

[4]In the massless case, the modular generators are ordinary differential operators, δ_0, of order 1. In the massive case, it has been conjectured that the modular generators are pseudo-differential operators $\delta_m = \delta_0 + \delta_r$, where the leading term is given by the massless generator δ_0 and δ_r is a pseudo-differential operators of order < 1. This second term is thought to give rise to non-local action without geometric interpretation.

the table.

She can hold out hope for a suitably general dynamical interpretation of modular automorphisms in a wide class of physically significant states. There is some indication that states of compact energy (e.g., states satisfying the Döplicher-Haag-Roberts and Buchholz-Fredenhagen selection criteria) give rise to well-behaved modular structure on wedges. In this case the wedge modular automorphisms can be related to those in the vacuum state by the Radon-Nikodym derivative (Borchers, 2000). The analogous problem for doublecones is still open.

Alternatively, she could reject the idea that the thermal time flow determines the temporal metric directly. Thermal time would only give rise to the order, topological, and group theoretic properties of physical time. Metrical properties would be determined by a completely different set of physical relations. Some support for this idea comes from the justification of the clock hypothesis in general relativity. Rather than stipulating the relationship between proper time, τ, and the length of a timelike curve $||\gamma||$, Fletcher (2013) shows that for any $\epsilon > 0$, there is an idealized lightclock moving along the curve which will measure $||\gamma||$ within ϵ. This justifies the clock hypothesis by linking the metrical properties of spacetime to the readings of tiny lightclocks. If the metrical properties of time experienced by localized observers arises via some physical mechanism akin to light clock synchronization. This would explain why the duration of time felt by the observer matches her proper time and not the geometrical flow of thermal time.

Perhaps motivated by the justification of the clock hypothesis, the defender of the TTH could attempt to argue that the metrical properties of time emerge from modular dynamics in the short distance limit of the theory. If the theory has a well-defined ultraviolet limit, the renormalization group flow should approach a conformal fixed point. Buchholz and Verch (1995) prove that in this limit, the double-cone modular operators act geometrically like wedge operators implementing proper time translations along the observer's worldline. It is unlikely that the physics at this scale would directly impact phenomenology, but the asymptotic connection might turn out to be important for explaining the metrical properties of spacetime (which bigger, more realistic lightclocks measure) as emergent features of some underlying theory of quantum gravity.

A final option would be to go back to the drawing board. Rovelli and Connes briefly note that since the modular automorphisms associated with each (faithful, normal) state of a von Neumann algebra are connected by

inner automorphisms, they all project down onto the same 1-parameter group of outer automorphisms the algebra. The TTH could be revised to claim that this canonical state-independent flow represents the non-metrical flow of physical time. It is not known, however, under what circumstances the outer flow acts in suitably geometric fashion to be interpretable as local dynamics, so it remains to be seen whether or not this is a viable option. The move does have immediate consequences for the global dynamics, however. Since the global algebra is expected to be type I, all modular automorphisms will be inner. As a result the canonical group of outer automorphisms is trivial. At a global level, there is no passage of time. At the local level, time emerges as a consequence of our ignorance of the global state.

4 The Classical Limit

The classical limit presents a different kind of challenge. Conceptually, nothing about the idea that a statistical state selects a preferred thermal time requires that the theory be quantum mechanical. The proposed mechanism for selecting a partial observable using modular theory, however, does appear to rely on the noncommutativity of quantum observables. If we model classical systems using abelian von Neumann algebras, then every state is tracial (i.e., $\phi(AB) = \phi(BA)$), and consequently every associated modular automorphism group acts as the identity, trivializing the thermal time flow. Does the TTH have a classical counterpart, or is quantum mechanics required to save time in a generally covariant setting?

Arguing by analogy with standard quantization procedures, Connes and Rovelli suggest that in the classical limit commutators need to be replaced by Poisson brackets. We begin with an arbitrary statistical state, ρ, represented by a probability distribution over a classical statespace Γ:

$$\int_{\Gamma} dx\ \rho(x) = 1 \tag{7}$$

where $x \in \Gamma$ is a timeless microstate. By analogy with the Gibbs postulate, we can introduce the "thermal Hamiltonian,"

$$H_{\rho} = -\ln \rho \tag{8}$$

With respect to the corresponding Hamiltonian vector field, the evolution of

an arbitrary classical observable, $f \in C^\infty(\Gamma)$, is given by

$$\frac{d}{ds} f = \{-\ln \rho, f\} \tag{9}$$

and $\rho = \exp(-H_\rho)$. With respect to the Poisson bracket structure, the classical algebra of observables is non-abelian. Gallavotti and Pulvirenti (1976) use this non-abelian structure to define an analogue of the KMS condition. Is this connection strong enough to support a version of the TTH in ordinary general relativity? Or does it only serve to aid us in understanding how the thermal time variable behaves in the transition from quantum theory to classical physics?

The difficulty lies in connecting the thermal time flow for an arbitrary statistical state to our ordinary conception of time. In the quantum case this link was provided by the Bisognano-Wichmann theorem, which does not have a classical analogue. The problem is magnified by the lack of a full understanding of statistical mechanics and thermodynamics in curved spacetime. Rovelli has done some preliminary work on developing a full theory of generally covariant thermodynamics based on the foundation supplied by the TTH, including an elegant derivation of the Tolman-Ehrenfest effect, but the field is still young.[5]

Setting aside these broader interpretive challenges for now, an important first step lies in obtaining a better understanding the classical selection procedure outlined above. As it turns out, the commutator-to-Poisson-bracket ansatz is on firmer foundational footing than one might initially suspect. As emphasized by Alfsen and Shultz (1998), non-abelian C^*-algebras have a natural *Lie-Jordan structure*:

$$AB = A \bullet B - i(A \star B) \ , \tag{10}$$

The non-associative Jordan product, $\bullet$, encodes information about the spectra of observables, while the associative Lie product, $\star$, encodes the generating relation between observables and symmetries. The significance of the commutator, is that it defines the canonical Lie product, $A \star B := i/2[A, B]$. Classical mechanical theories formulated on either a symplectic or Poisson manifold have a natural Lie-Jordan structure as well. The standard product of functions defines an associative Jordan product, encoding spectral information, while the Poisson bracket determines the associative Lie product,

[5]See Rovelli and Smerlak (2011).

describing how classical observables generate Hamiltonian vector fields on statespace. Together, this structure is called a *Poisson algebra*. The primary difference between the classical and quantum cases is the associativity/non-associativity of the Jordan product.

These considerations point towards the idea that the appropriate classical analogue of a noncommutative von Neumann algebra, is not a commutative von Neumann algebra, but a Poisson algebra. In this setting, initial strides towards a classical analogue of modular theory have been made by Weinstein (1997). Given any smooth density, μ, on a Possion manifold, Γ, Weinstein defines a corresponding *modular vector field* ϕ_μ given by the operator $\phi_\mu : f \to \mathrm{div}_\mu H_f$ where H_f is the Hamiltonian vector field associated with a classical observable, $f \in C^\infty(\Gamma)$. The antisymmetry of the Poisson bracket entails that the operator ϕ_μ is a vector field on Γ. Weinstein proposes ϕ_μ as the classical analogue of the modular automorphism group. It characterizes the extent to which the Hamiltonian vector fields are divergence free (with respect to the density μ), vanishing iff all Hamiltonian vector fields are divergence free.

We can connect Weinstein's classical modular theory to the TTH. If Γ is a symplectic manifold and we let μ be the density associated with the canonical Liouville volume form, then $\phi_\mu(f) = 0$ for all observables. This reflects the conservation of energy by Hamiltonian flows in symplectic dynamical systems. Given any statistical state, however, we can define an associated density which leads to a nontrivial modular vector field. For any positive function, h, we have

$$\phi_{h\mu} = \phi_\mu + H_{-\ln h} = H_{-\ln h}. \tag{11}$$

Therefore any statistical state, ρ, defines a modular vector field equivalent to the Hamiltonian vector field $H_{-\ln \rho}$ associated with the density $e^{-\ln \rho}\mu$. We immediately recognize $-\ln \rho$ as the thermal Hamiltonian postulated by Connes and Rovelli. Clearly, $e^{is \ln \rho} \rho e^{-is \ln \rho} = \rho$, thus the state is invariant with respect to the flow of $H_{-\ln \rho}$. Additionally, it can be shown that ρ satisfies the KMS condition with respect to these dynamics, hence, from the perspective of the associated time flow ρ resembles an invariant equilibrium state just as in the quantum case.

5 Conceptual Challenges

As we have seen in the previous two sections, the TTH faces a number of technical challenges (some of which look easier to overcome than others). There are, however, several deeper conceptual problems looming in the background which pose a more serious challenge to the viability of the hypothesis. Here, we will discuss two of the most pressing.

The first, which we will call the *generality problem*, draws upon the preceding discussion of the classical limit. While mathematically speaking, Weinstein's modular vector field gives us a method for selecting a canonical thermal time flow in a classical theory, physical speaking, there is no reason why we should view the corresponding thermal time as physical time. As we have seen, any statistical state determines thermal dynamics according to which it is a KMS state, however, if ρ is a non-equilibrium state, the resultant thermal time flow does not align with our ordinary conception of time. By the lights of thermal time, a cube of ice in a cup of hot coffee is an invariant equilibrium state! The same problem arises in the quantum domain — only for states which are true equilibrium states will the thermal time correspond to physical time.

It appears inevitable that the TTH will have to be tempered. Rather than letting any state determine a corresponding flow of thermal time, only certain reference states should be permitted. Apart from the problem of providing an intrinsic, non-dynamical characterization of such states, if a system is not in one of these, it is hard to envision how a counterfactual state of affairs can determine the actual flow of time.[6] This might provide more reasons for the defender of the TTH to explore the state independent, outer modular flow. Alternatively, she could try to argue that local non-equilibrium behavior can be viewed as small fluctuations from some background state. On this approach, the local flow of time in my office according to which the ice

[6] A closely related worry, what we might call the *background-dependence problem*, has been voiced by Earman (2011) and Ruetsche (2014). Their concern is that we can only identify modular automorphisms as dynamics because we already have a rich spatiotemporal geometry in the background. This casts doubt on whether the TTH can provide a coherent definition of time in situations where such structure is absent (as required to solve the full problem of time). This is exacerbated if the TTH is modified in response to the generality problem. Unless the modular automorphism group can always be viewed dynamically, the defender of the TTH will be hard pressed to find constraints capable of separating the dynamical cases from the non-dynamical cases which are independent of all background temporal structure.

melts and the coffee cools is not defined by the thermal state of the ice/coffee system, but the thermal state of some larger enveloping system (the entire universe perhaps). Rovelli (1993) hints in this direction, calculating that in a Friedman-Robertson-Walker universe, the thermal time induced by the equilibrium state of the cosmic microwave background will be proportional to the FRW time. While the connection is intriguing, it seems unlikely that an explanation of this sort will be able to account for the flow of time experienced by localized, mortal observers like us. It would be truly remarkable to discover that our faculties of perception are sensitive to the thermal features of the CMB.

The second problem is the *gauge problem*. The TTH does succeed in providing a means to select a privileged 1-parameter flow on the space of full, gauge invariant observables of a generally covariant theory. What makes this flow interpretable as a *dynamical* flow, however, is its description as a sequence of correlations between partial observables. The difficulty is that these partial observables are not diffeomorphism invariant. Assuming that we treat diffeomorphisms in generally covariant theories as standard gauge symmetries (which is how we got into the problem of time in the first place), then the partial observables are just descriptive fluff. They do not directly represent physical features of our world.

The problem is *not* the resultant timelessness of fundamental physics. The TTH adopts this dramatic conclusion willingly. The problem is that the TTH is supposed to explain how the appearance of time and change emerge from timeless foundations. But the explanation given is couched in gauge-dependent language, and it is not apparent how we can extract a gauge invariant story from it. We can introduce partial observables and use correlations between them to calculate and predict emergent dynamical behavior, but we cannot use these correlations to *explain* that behavior. We lack a gauge invariant picture of generally covariant theories, and the TTH, at least in its present form, does not provide one.

Can a revised TTH give us the explanatory tools needed to understand the flow of time without reference to partial observables, or, does the entire framework of timeless mechanics require us to revise our conception of how ontology, explanation, and gauge symmetries are related?[7] Whether or not

[7]Drifting in the latter direction, Rovelli (2014) suggests that gauge-dependent quantities are more than just mathematical redundancies, "they describe handles through which systems couple: they represent real relational structures to which the experimentalist has access in measurement by supplying one of the relata in the measurement procedure itself."

quantum thermodynamics can save time may rest on the solutions to these new incarnations of vexingly familiar philosophical problems.

References

Alfsen, E. and F. Shultz (1998). Orientation in operator algebras. *Proceedings of the National Academy of Sciences, USA 95*, 6596–6601.

Borchers, H. J. (2000). On revolutionizing quantum field theory with Tomita's modular theory. *Journal of Mathematical Physics 41*(6), 3604–3673.

Brunetti, R., K. Fredenhagen, and R. Verch (2003). The generally covariant locality principle – a new paradigm for local quantum field theory. *Communications in Mathematical Physics 237*, 31–68.

Buchholz, D. and R. Verch (1995). Scaling algebras and renormalization group in algebraic quantum field theory. *Reviews in Mathematical Physics 7*, 1195.

Connes, A. and C. Rovelli (1994). Von Neumann algebra automorphisms and time-thermodynamics relation in generally covariant quantum theories. *Classical and Quantum Gravity 11*(12), 2899.

Earman, J. (2002). Thoroughly modern McTaggart. *Philosopher's Imprint*, 2. http://www.philosophersimprint.org/002003/.

Earman, J. (2011). The Unruh effect for philosophers. *Studies in History and Philosophy of Modern Physics 42*, 81–97.

Fletcher, S. (2013). Light clocks and the clock hypothesis. *Foundations of Physics 43*, 1369–1383.

Gallavotti, G. and M. Pulvirenti (1976). Classical KMS condition and Tomita-Takesaki theory. *Communications in Mathematical Physics 46*, 1–9.

Hislop, P. D. and R. Longo (1982). Modular structure of the local algebras associated with a free massless scalar field theory. *Communications in Mathematical Physics 84*, 71.

Jian-yang, Z., B. Aidong, and Z. Zheng (1995). Rindler effect for a nonuniformly accelerating observer. *International Journal of Theoretical Physics 34*, 2049–2059.

Martinetti, P. and C. Rovelli (2003). Diamond's temperature: Unruh effect for bounded trajectories and thermal time hypothesis. *Classical and Quantum Gravity 20*(22), 4919.

Paetz, T.-T. (2010). An analysis of the 'thermal-time concept' of Connes and Rovelli. Master's thesis, Georg-August-Universität Göttingen.

Rovelli, C. (1993). The statistical state of the universe. *Class.Quant.Grav. 10*, 1567.

Rovelli, C. (2011). Forget time: Essay written for the FQXi contest on the nature of time. *Foundations of Physics*.

Rovelli, C. (2014). Why gauge? *Foundations of Physics 44*(1), 91–104.

Rovelli, C. and M. Smerlak (2011). Thermal time and Tolman–Ehrenfest effect: 'temperature as the speed of time'. *Classical and Quantum Gravity 28*(7), 075007.

Ruetsche, L. (2014). Warming up to thermal the thermal time hypothesis. Quantum Time Conference, University of Pittsburgh, March 28-29.

Saffary, T. (2005). *Modular Action on the Massive Algebra*. Ph. D. thesis, Hamburg.

Trebels, S. (1997). *Über die Geometrische Wirkung Modularer Automoprhismen*. Ph. D. thesis, Göttingen.

Weinstein, A. (1997). The modular automorphism group of a Poisson manifold. *Journal of Geometry and Physics 23*, 379–394.

Neural redundancy and its relation to neural reuse

Abstract

Evidence of the pervasiveness of neural reuse in the human brain has forced a revision of the standard conception of modularity in the cognitive sciences. One persistent line of argument against such revision, however, draws from a large body of experimental literature attesting to the existence of cognitive dissociations. While numerous rejoinders to this argument have been offered over the years, few have grappled seriously with the phenomenon. This paper offers a fresh perspective. It takes the dissociations seriously, on the one hand, while affirming that traditional modularities of mind do not do justice to the evidence of neural reuse, on the other. The key to the puzzle is neural redundancy. The paper offers both a philosophical analysis of the relation between reuse and redundancy, as well as a plausible solution to the problem of dissociations.

1. Introduction

Cognitive science, linguistics and the philosophy of psychology have long been under the spell of "the modularity of mind" (Fodor 1983), or the idea of the mind as a modular system (see e.g. de Almeida and Gleitman 2018). In contemporary psychology, a modular system is generally understood to be "one consisting of functionally specialized subsystems responsible for processing different classes of input (e.g. for vision, hearing, human faces, etc.), or at any rate for handling specific cognitive tasks" (Zerilli 2017a, 231). According to this theory, "human cognition can be decomposed into a number of functionally independent processes, [where] each of these processes operates over a distinct domain of cognitive information" (Bergeron 2007, 176). What makes one process distinguishable from another is its "functional independence, the fact that one can be affected, in part or in totality, without the other being affected, and vice versa" (Bergeron 2007, 176). Furthermore, given that functional processes are realized in the brain, a functionally specialized process is one which presumably occupies a distinctive portion of neural tissue, though not necessarily a small, closely circumscribed and contiguous region. So fruitful and influential has this model been that it is safe to say that in many quarters of the cognitive sciences—and most especially in cognitive psychology, cognitive neuropsychology and evolutionary psychology—modularity is essentially the received view (McGeer 2007; Carruthers 2006; de Almeida and Gleitman 2018).

Developments in cognitive neuroscience over the past thirty years, however, have discomfited the modular account. More evidence than ever before points to the pervasiveness of neural reuse in the human brain—the "redeployment" or "recycling" of neural circuits over widely disparate cognitive domains (Anderson, 2010, 2014; Dehaene, 2005). As the terminology suggests, theories of "re-use" posit the "exaptation" of established and diachronically stable neural circuits over the course of evolution or normal development *without* loss of original function, so that the functional contribution of a circuit is preserved across multiple task domains.[1] As Anderson (2010, 246) explains, "rather than posit a functional architecture for the brain whereby individual regions are dedicated to large-scale cognitive domains like vision, audition, language and the like, neural reuse theories suggest that low-level neural circuits are used and reused for various purposes in different cognitive and task domains." According to the theory, just the same circuits exapted for one purpose can be exapted for another provided sufficient intercircuit pathways exist to allow alternative arrangements of them. Indeed, the same parts put together in *different* ways will yield different functional outcomes, just as "if one puts together the same parts *in the same way* one will get the same functional outcomes" (Anderson 2010, 247, my emphasis). The evidence here converges from heterogeneous sources and research paradigms, including neuroimaging (Anderson 2007a; 2007b; 2007c; 2008), computational (Eliasmith 2015), biobehavioral (Casasanto and Dijkstra 2010) and interference paradigms (Gauthier et al.

[1] This usage of "exaptation" is somewhat misleading, since exaptation usually implies loss of original function (see Godfrey-Smith 2001).

2003), and exempts practically no area of the brain (Leo et al. 2012, 2), including areas long regarded as specialized hubs for certain types of sensory processing, e.g. visual and auditory pathways (Striem-Amit and Amedi 2014). Among other things, this means that one of the hallmark features of a module—its domain specificity (Coltheart 1999)—looks too stringent a requirement to prove useful.[2] For neural reuse demonstrates that any one module will typically be sensitive to *more* than one stimulus, including—most importantly—those channeled along intermodal pathways. Meanwhile efforts to salvage a computational or "software" theory of modularity, which carries no commitments regarding implementation, have met with scepticism (Anderson 2007c; 2010; Anderson & Finlay 2014) if not outright opposition (Zerilli 2017a).[3] And while the brain could still be modular in some other sense, what is clear is that the strict domain-specific variety of modularity can no longer serve as an appropriate benchmark.[4]

And yet there is a persistent line of argument *against* this conclusion which draws from a large body of experimental literature attesting to the existence of cognitive

[2] The sense of domain specificity that is relevant here refers to a module's sensitivity to a restricted class of inputs as defined by a domain of psychology—such as visual, auditory or linguistic information. For discussion of alternative senses, see Barrett and Kurzban (2006) and Prinz (2006).

[3] Though by no means universally (see e.g. Carruthers 2010; Jungé and Dennett 2010).

[4] Nor, for that matter, can its cognate property, informational encapsulation (see below).

dissociations, in which a cognitive ability (say language) is either selectively impaired (linguistic ability is compromised, but no other cognitive ability seems to be materially affected) or selectively spared (general intelligence is compromised, while linguistic abilities function more or less as they should). This literature, most vividly exemplified in lesion studies, is frequently cited in support of classical modularities of mind—be they inspired by the likes of Jerry Fodor (1983), evolutionary psychology (e.g. Cosmides and Tooby 1994; Barrett and Kurzban 2006; Carruthers 2006) or some variation thereof (e.g. ACT-R). While numerous rejoinders to this line of thinking have been offered over the years, few have grappled seriously with the phenomenon, either dismissing the dissociations as noisy, or reasoning from architectural considerations that even nonmodular systems can generate dissociations (Plaut 1995). The aim of this paper is to offer a fresh perspective on this vexed topic. I take the dissociation evidence seriously, on the one hand, while affirming that traditional modularities of mind do not do justice to the evidence of neural reuse, on the other. I do this by invoking neural redundancy, an important feature of cortical design that ensures we have various copies of the same elementary processing units that can be put to alternative (if computationally related) uses in enabling diverse cognitive functions. In the course of the discussion I offer a philosophical explication of the relationship between neural reuse and neural redundancy.

2. What is the Problem? Cognitive Dissociations and Neural Reuse

Let us take an especially contentious question to underscore the nature of the problem we are dealing with and how redundancy might assist in its illumination. The question is this: Does language rely on specialized cognitive and neural machinery, or does it rely on the same machinery that allows us to get by in other domains of human endeavour? The question is bound up with many other questions of no less importance, questions concerning the uniqueness of the human mind, the course of biological evolution and the power of human culture. What is perhaps a little unusual about this question, however—unusual for a question whose answer concerns both those working in the sciences and the humanities—is that it can be phrased as a polar interrogative, i.e. as a question which admits of a yes or no response. And indeed the question has divided psychologists, linguists and the cognitive science community generally for many decades now, more or less into two camps. I would like to sketch the beginnings of an answer to this question—and others like it—in a way that does not pretend it can receive a simple yes or no response.

First of all, let me stress again that neural reuse is as well verified a phenomenon as one can expect in the cognitive sciences, and that it has left virtually no domain of psychology untouched. Neural reuse suggests that there is nothing so specialized in the cortex that it cannot be repurposed to meet new challenges while retaining its capacity for meeting old ones. In that regard, to be sure, what I am proposing is unapologetically on the side of those who maintain that language, as well as many other psychological capacities, are

not cognitively special—e.g. that there is no domain-specific "language organ" (cf. Chomsky 1980,39, 44; 1988, 159; 2002, 84-86).

And yet I would like to carefully distinguish this claim from the claim that there are no areas of the brain that subserve exclusively linguistic functions. The neuropsychological literature offers striking examples of what appear to be fairly clean dissociations between linguistic and nonlinguistic capacities, i.e. cases in which language processing capacities appear to be disrupted without impeding other cognitive abilities, and cases in which the reverse situation holds (Fedorenko et al. 2011; Hickok and Poeppel 2000; Poeppel 2001; Varley et al. 2005; Luria et al. 1965; Peretz and Coltheart 2003; Apperly et al. 2006). An example would be where the ability to hear words is disrupted, but the ability to recognize non-word sounds is spared (Hickok and Poeppel 2000; Poeppel 2001). Discussing such cases, Pinker and Jackendoff (2005, 207) add that "[c]ases of amusia and auditory agnosia, in which patients can understand speech yet fail to appreciate music or recognize environmental sounds…show that speech and non-speech perception in fact doubly dissociate." Although dissociations are to some extent compatible with reuse—indeed there is work suggesting that focal lesions can produce specific cognitive impairments within a range of nonclassical architectures (Plaut 1995)—and it is equally true that often the dissociations reported are noisy (Cowie 2008), still their very ubiquity needs to be taken seriously and accounted for in a more systematic fashion than many defenders of reuse have been willing to do (see e.g. Anderson 2010, 248; 2014, 46-48). After all, a good deal of support for

theories of reuse comes from the neuroimaging literature, which is somewhat ambiguous taken by itself. As Fedorenko et al. (2011, 16428) explain:

> standard functional MRI group analysis methods can be deceptive: two different mental functions that activate neighbouring but non-overlapping cortical regions in every subject individually can produce overlapping activations in a group analysis, because the precise locations of these regions vary across subjects, smearing the group activations. Definitively addressing the question of neural overlap between linguistic and nonlinguistic functions requires examining overlap within individual subjects, a data analysis strategy that has almost never been applied in neuroimaging investigations of high-level linguistic processing.

When Fedorenko and her colleagues applied this strategy themselves, they found that "most of the key cortical regions engaged in high-level linguistic processing are not engaged by mental arithmetic, general working memory, cognitive control or musical processing," and they think that this indicates "a high degree of functional specificity in the brain regions that support language" (2011, 16431). While I do not believe that claims of this strength have the least warrant—as I shall explain, functional specificity cannot be established merely by demonstrating that a region is selectively engaged by a task—these results do at least substantiate the dissociation literature in an interesting way and make it more difficult for

those who would prefer to dismiss the dissociations with a ready-made list of alternative explanations. Similar results were found by Fedorenko et al. (2012).

3. How Might Redundancy Feature In a Solution?

With rare exceptions (e.g. Friston and Price 2003; Barrett and Kurzban 2006; Jungé and Dennett 2010), redundancy has passed almost unnoticed in the philosophical and cognitive science literature. This is in stark contrast to the epigenetics literature, where redundancy and the related concept of degeneracy[5] have been explored to some depth (e.g. see Edelman and Gally 2001; Mason 2010; Whiteacre 2010; Deacon 2010; Iriki and Taoka 2012; Maleszka et al. 2013). The idea behind neural redundancy is that, for good evolutionary reasons (see below), the brain incorporates a large measure of redundancy of function. Brain regions (such as cortical columns and similar structures) fall in an iterative, repetitive and almost lattice-like arrangement in the cortex. Neighbouring columns have similar response properties: laminar and columnar changes are for the most part smooth—not abrupt—as one moves across the cortex, and adjacent modules do not differ markedly from one another in their basic structure and computations (if they really differ at all when taken in such

[5] Redundancy occurs when items have the same structure and function (i.e. are both isomorphic and isofunctional). Degeneracy occurs when items having *different* structures can perform the same function (i.e. are heteromorphic but isofunctional). Degeneracy implies genuine multiple realization (see Zerilli 2017b).

proximity). Regional *solitarity* is therefore not likely to be a characteristic of the brain (Anderson 2014, 141).[6] That is to say, we do not possess just one module for X, and one module for Y, but in effect several *copies* of the module for X, and several copies of the module for Y, all densely stuffed into the same cortical zones. As Buxhoeveden and Casanova (2002, 943) explain of neurons generally:

> In the cortex, more cells do the job that fewer do in other regions....As brain evolution paralleled the increase in cell number, a reduction occurred in the sovereignty of individual neurones; fewer of them occupy critical positions. As a consequence, plasticity and redundancy have increased. In nervous systems containing only a few hundred thousand neurones, each cell plays a more essential role in the function of the organism than systems containing billions of neurones.

The same principle very likely holds for functionally distinct groupings of neurons (i.e. cortical columns and like structures), as Jungé and Dennett (2010, 278) conjecture:

> It is possible that specialized brain areas contain a large amount of structural/computational redundancy (i.e., many neurons or collections of neurons

[6] The term "solitarity" is Anderson's, but while he concedes that solitarity will be "relatively rare," he does not appear to believe that anything particularly significant follows from this. See also Anderson (2010, 296).

> that can potentially perform the same class of functions). Rather than a single neuron or small neural tract playing roles in many high-level processes, it is possible that distinct subsets of neurons within a specialized area have similar competencies, and hence are redundant, but as a result are available to be assigned individually to specific uses....In a coarse enough grain, this neural model would look exactly like multi-use (or reuse).

This is plausibly why capacities which are functionally very closely related, but which for whatever reason are forced to recruit different neural circuits, will often be localized in broadly the same regions of the brain. For instance, first and second languages acquired early in ontogeny settle down in nearly the same region of Broca's area; and even when the second language is acquired in adulthood the second language is represented nearby within Broca's area (while artificial languages are not) (Kandel & Hudspeth 2013). The neural coactivation graphs of such composite networks must look very similar. Indeed these results suggest—and a redundancy model would predict—that two very similar tasks which are forced to recruit different neural circuits should exhibit similar patterns of activation. And this is more or less what we find (see below).

One might be tempted to think that redundancy and reuse pull in opposite directions. This is because whereas reuse posits that neural circuits get reused across different tasks and task categories, redundancy accommodates the likelihood of diverse

cognitive functions being activated by structurally and computationally equivalent circuits running in parallel: instead of a single circuit being reused across domains, two, three or more *copies* of that same circuit may be recruited differentially across those domains, such that no *single* circuit gets literally "re-used." But there is no substantive tension here. The redundancy account in truth *supplements* the reuse picture in a way that is consistent with the neuroimaging data, faithful to the core principle of reuse, and compatible with the apparent modularization and separate modifiability of technical and acquired skills in ontogeny. Evidence of the reuse of neural circuits to accomplish different tasks has, in fact, been adduced in aid of a theory which posits the reuse of the same neural *tokens* to accomplish these different tasks. Redundancy means we must accept that at least some of the time what we may actually be witnessing is reuse of the same *types* to accomplish these tasks. This does not diminish the standing of reuse. Let me explain.[7]

To the extent that a particular composite reuses types, and is dissociable pro tanto—residing in segregated brain tissue that is not active outside the domain in question—it is true that to that extent its constituents will *appear* to be domain-specific. But in this case looks will be deceiving. The classical understanding of domain specificity in effect *assumes* solitarity—that a module for X does something which no other module can do as well, or

[7] For a developmental twist on the type/token distinction invoked in the context of modular theorizing about the mind, see Barrett (2006).

that even *if* another module can do X as well, taken together these X-ing modules do not perform outside the X-domain. Here is an example of the latter idea (Bergeron 2007, 176):

> a pocket calculator could have four different division modules, one for dividing numbers smaller than or equal to 99 by numbers smaller than or equal to 99, a second one for dividing numbers smaller than or equal to 99 by numbers greater than 99, a third one for dividing numbers greater than 99 by numbers greater than 99, and a fourth one for dividing numbers greater than 99 by numbers smaller than or equal to 99. In such a calculator, these four capacities could all depend on (four versions of) the same algorithm. Yet, random damage to one or more of these modules in a number of such calculators could lead to observable (double) dissociations between any two of these functions.

Here, each module performs fundamentally the same algorithm, but in distinct hardware, such that dissociations are observable between any two functions. Notice, however, that none of these modules performs outside the "division" domain. This is what allows such duplicate modules to be considered domain-specific—they perform functions which, for all that they might run in parallel on duplicate hardware, are unique to a specific domain of operation, in this case division. If such modules could do work outside the division domain, they would lose the status of domain specificity, and acquire the status of domain neutrality (i.e. they would be domain-general). This is why a module that appears dedicated to a

particular function may not be domain-specific in the classical sense. Dedication is not the same as domain specificity, and redundancy, whether of calculator algorithms or neural circuits, explains why. A composite of neural regions will be dedicated without being domain-specific if its functional resources are accessible to other domains through the deployment (reuse) of neural surrogates (i.e. redundant or "proxy" tokens). In this case its constituents will be multi-potential but single-use (Jungé & Dennett 2010, 278), and the domain specificity on display somewhat cosmetic. To take an example with more immediate relevance to the brain, a set of cortical columns that are structurally and computationally similar may be equally suited for face recognition tasks, abstract-object recognition tasks, the recognition of moving objects, and so on. One of these columns could be reserved for faces, another for abstract objects, another for moving objects, and so on. What is noteworthy is that while the functional activation may be indistinguishable in each case, and the same *type* of resource will be employed on each occasion, a different *token* module will be at work at any one time. To quote Jungé and Dennett (2010, 278) again:

> In an adult brain, a given neuron [or set of neurons] would be aligned with only a single high-level function, whereas each area of neurons would be aligned with very many different functions.

Such modules (and composites) are for all intents and purposes *qualitatively* identical, though clearly not *numerically* identical, meaning that while they share their properties, they

are not *one and the same* (Parfit 1984). The evidence of reuse is virtually all one way when it comes to the pervasiveness of functional inheritance across cognitive domains. It may be that this inheritance owes to reuse of the same tokens (literal reuse) or to reuse of the same types (reuse by proxy), but the inheritance itself has been amply attested. This broader notion of reuse still offers a crucial insight into the operations of cognition, and I dare say represents a large part of the appeal of the original massive redeployment hypothesis (Anderson 2007c).

It is interesting to note in this respect that although detractors have frequently pointed out the ambiguity of neuroimaging evidence on account of its allegedly coarse spatial resolution (e.g. Carruthers 2010), suggesting that the same area will be active across separate tasks and task categories even if distinct but spatially adjacent and/or interdigitated circuits are involved in each case, this complaint can have no bearing on reuse by proxy. Fedorenko et al. (2011, 16431) take their neuroimaging evidence to support "a high degree of functional specificity in the brain regions that support language," but their results do not license this extreme claim. The regions they found to have been selectively engaged by linguistic tasks were all adjacent to the regions engaged in nonlinguistic tasks. Elementary considerations suggest that they have discovered a case of reuse by proxy involving language: the domains tested (mental arithmetic, general working memory, cognitive control and musical processing) make use of many of the same computations as high-level linguistic processing, even though they run them on duplicate hardware. Redundancy makes it is easy to see how fairly sharp dissociations could arise—knocking out one token module need

disrupt only one high-level operation: other high-level operations that draw on the same *type* of resource may well be spared.

The consequences of this distinction between literal reuse and reuse by proxy for much speculation about the localization and specialization of function are potentially profound. In cognitive neuropsychology the discovery that a focal lesion selectively impairs a particular cognitive function is routinely taken as evidence of its functional specificity (Coltheart 2011; Sternberg 2011). Even cognitive scientists who take a developmental approach to modularity, i.e. who concede that parts of the mind may be modular but stress that modularization is a developmental process, concede too much when they imply, as they frequently do, that modularization results in domain-specific modules (Karmiloff-Smith 1992; Prinz 2006; Barrett 2006; Cowie 2008; Guida et al. 2016). This is true in some sense, but not in anything like the standard sense, for redundancy envisages that developmental modules form a special class of neural networks, namely those which are *qualitatively* identical but *numerically* distinct. The appearance of modularization in development is thus fully compatible with deep domain interpenetration. In any event redundancy does not predict that all acquired skills will be modular. The evidence suggests that while some complex skills reside in at least partly dissociable circuitry, most complex skills are implemented in more typical neural networks, i.e. those consisting of literally shared parts.[8]

[8] This seems to be true regardless of whether the complex skills are innate or acquired.

4. What Else Might Redundancy Explain?

It is generally a good design feature of any system to have spare capacity. For instance, in engineered systems, "redundant parts can substitute for others that malfunction or fail, or augment output when demand for a particular output increases" (Whiteacre 2010, 14). The positive connection between robustness and redundancy in biological systems is also clear (Edelman and Gally 2001; Mason 2010; Whiteacre 2010; Iriki & Taoka 2012). So there are good reasons for evolution to have seen to it that our brains have spare capacity. But in the case of the brain and the cortex most especially, there are other reasons why redundancy would be an important design feature. It offers a solution to what Jungé and Dennett (2010, 278) called the "time-sharing" problem. It may also offer a solution to what I call the "encapsulation" problem.

The time-sharing problem arises when multiple simultaneous demands are made on the same cognitive resource. This is probably a regular occurrence. Here are just a few examples.

- Driving a car and holding a conversation at the same time: if it is true that some of the selfsame motor operations underlying aspects of speech production and comprehension are also required for the execution of sequenced or complex motor functions (Pulvermüller and Fadiga 2010; Graziano et al. 2002; MacNeilage 1998; Glenberg et al.

2008; Glenberg and Kaschak 2002; Glenberg et al. 2007; Greenfield 1991), as perhaps exemplified by driving a manual vehicle or operating complex machinery (e.g. playing the organ), how do we manage to pull this off?

- By reflecting the recursive structure of thought (Christiansen and Chater 2016, 51), the language circuits may redeploy a recursive operation simultaneously during sentence production. This might be the case during the formation of an embedded relative clause—the thought and its encoding may require parallel use of the same sequencing principle. Again, how do we manage this feat?
- If metarepresentational operations are involved in the internalization of conventional sound-meaning pairs, and also in the pragmatics and mindreading that carry on simultaneously during conversation, as argued by Suddendorf (2013), could this not simply be another instance of time-sharing? The example is contentious, but it still raises the question: how does our brain manage to do things like this?
- Christiansen and Chater's (2016) "Chunk and Pass" model of language processing envisages *multilevel* and *simultaneous* chunking procedures. As they put it, "the challenge of language acquisition is to learn a dazzling sequence of rapid processing operations" (2016, 116). What must the brain be like to allow for this dazzling display?

Explaining these phenomena is difficult. Indeed when dealing with clear (literal) instances of reuse, results from the interference paradigm show that processing bottlenecks are inevitable—true multi-tasking is impossible. Redundancy offers a natural explanation of how

the brain overcomes the time-sharing problem. It explains, in short, how we are able to walk and chew gum at the same time.

Redundancy might also offer a solution to what I have called the encapsulation problem. The neural networks that implement cognitive functions are not likely to be characterized by informational encapsulation if they share their nodes with networks implementing other cognitive functions. This is because in sharing their nodes with these other systems they will prima facie have access to the information stored and manipulated by those other systems (Anderson 2010, 300). If, then, overlapping brain networks must share information (Pessoa 2016, 23), it would be reasonable to suppose that central and peripheral systems do *not* overlap. For peripheral systems, which are paradigmatically fast and automatic, would not be able to process inputs as efficiently if there were a serious risk of central system override—i.e. of beliefs and other central information getting in the way of automatic processing. But we know from the neuroimaging literature that quite often the brain networks implementing central and peripheral functions *do* overlap. This is puzzling in light of the degree of cognitive impenetrability that certain sensory systems still seem to exhibit—limited though it may be. If it is plausible to suppose that the phenomenon calls for segregated circuitry, redundancy could feature in a solution to the puzzle, since it naturally explains how the brain can make parallel use of the same resources. Neuroimaging maps might well display what appear to be overlapping brain regions between two tasks (one involving central information, the other involving classically peripheral operations), but the

overlap would not exist—there would be distinct albeit adjacent or interdigitated and nearly identical circuits recruited in each case. Of course there may be other ways around the encapsulation problem that do not require segregated circuitry: the nature and extent of the overlap is presumably important. But clearly redundancy opens up some fascinating explanatory possibilities.

To the extent that acquired skills must overcome both the time-sharing problem as well as the encapsulation problem—for acquired competencies are often able to run autonomously of central processes—we might expect that their neural implementations incorporate redundant tissue. In concluding, let me illustrate this point by offering a gloss on a particular account of how skills and expertise are acquired during development elaborated by Guida et al. (2016) and Anderson (2014). The process involved is called "search" (Anderson 2014). Search is an exploratory synaptogenetic process, "the active testing of multiple neuronal combinations until finding the most appropriate one for a specific skill, i.e., the neural niche of that skill" (Guido et al. 2016, 13). The theory holds that in the early stages of skill acquisition, the brain must search for an appropriate mix of brain areas, and does so by recruiting relatively widely across the cortex. When expertise has finally developed, a much narrower and more specific network of brain areas has been settled upon, such that "[a]s a consequence of their extended practice, experts develop domain-specific knowledge structures" (Guido et al. 2016, 13). The gloss (and my hunch) is this: first, that repeated practice of a task that requires segregation (to get around time-

sharing and encapsulation issues) will in effect *force* search into redundant neural territory (Karmiloff-Smith 1992; Barrett 2006; Barret and Kurzban 2006); second, that search will recruit idle or relatively underutilized circuits in preference to busy ones as a general default strategy. Guido et al. (2016) cite evidence that experts' brains reuse areas for which novices' brains make only limited use: "Whereas novices use episodic long-term memory areas (e.g., the mediotemporal lobe) for performing long-term memory tasks, experts are able to (re)use these areas also for performing working-memory tasks" (Guido et al. 2016, 14). Guido and colleagues, in agreement with Anderson (2014), seem to have literal reuse in mind. But the same evidence they cite is consistent with reuse by proxy. As Barrett and Kurzban (2006, 639) suggest, echoing a similar suggestion by Karmiloff-Smith (1992), a developmental system

> could contain a procedure or mechanism that partitioned off certain tasks—shunting them into a dedicated developmental pathway—under certain conditions, for example, when the cue structure of repeated instances of the task clustered tightly together, and when it was encountered repeatedly, as when highly practiced....Under this scenario, reading could still be recruiting an evolved system for object recognition, and yet phenotypically there could be *distinct modules* for reading and for other types of object recognition.

5. Conclusion

It is true that language and other cognitive skills frequently dissociate from other skills, but redundancy puts this sort of modularization in its proper context. Redundancy predicates functional inheritance across tasks and task categories even when the tasks are implemented in spatially segregated neural networks. Thus dissociation evidence alone does not always indicate true functional specificity. In particular, these dissociations provide no evidence that language is cognitively special vis-à-vis other cognitive domains.

References

Anderson, Michael L. 2007a. "Evolution of Cognitive Function via Redeployment of Brain Areas." *The Neuroscientist* 13:13-21.

——2007b. "Massive Redeployment, Exaptation, and the Functional Integration of Cognitive Operations." *Synthese* 159 (3): 329-345.

——2007c. "The Massive Redeployment Hypothesis and the Functional Topography of the Brain." *Philosophical Psychology* 21 (2): 143-174.

——2008. "Circuit Sharing and the Implementation of Intelligent Systems." *Connection Science* 20 (4): 239-251.

——2010. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences* 33 (4): 245-266; discussion 266-313.

——2014. *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Anderson, Michael L., and Barbara L. Finlay. 2014. "Allocating Structure to Function: The Strong Links Between Neuroplasticity and Natural Selection." *Frontiers in Human Neuroscience* 7:1-16.

Apperly, I.A., D. Samson, N. Carroll, S. Hussain, and G. Humphreys. 2006. "Intact First- and Second-Order False Belief Reasoning in a Patient with Severely Impaired Grammar." *Social Neuroscience* 1 (3-4): 334-348.

Barrett, H. Clark. 2006. "Modularity and Design Reincarnation." In *The Innate Mind Volume 2: Culture and Cognition*, ed. Peter Carruthers, Stephen Laurence, and Stephen P. Stich, 199-217. New York: Oxford University Press.

Barrett, H. Clark, and Robert Kurzban. 2006. "Modularity in Cognition: Framing the Debate." *Psychological Review* 113 (3): 628-647.

Bergeron, Vincent. 2007. "Anatomical and Functional Modularity in Cognitive Science: Shifting the Focus." *Philosophical Psychology* 20 (2): 175-195.

Buxhoeveden, Daniel P., and Manuel F. Casanova. 2002. "The Minicolumn Hypothesis in Neuroscience." *Brain* 125:935-951.

Carruthers, Peter. 2006. The *Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Casasanto, D., and K. Dijkstra. 2010. "Motor Action and Emotional Memory." *Cognition* 115 (1): 179-185.

Chomsky, Noam. 1980. *Rules and Representations*. New York: Columbia University Press.

——1988. *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.

——2002. *On Nature and Language*. New York: Cambridge University Press.

Christiansen, Morten H., and Nick Chater. 2016. *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge, MA: MIT Press.

Coltheart, Max. 1999. "Modularity and Cognition." *Trends in Cognitive Sciences* 3 (3): 115-120.

——2011. "Methods for Modular Modelling: Additive Factors and Cognitive Neuropsychology." *Cognitive Neuropsychology* 28 (3-4): 224-240.

Cosmides, Leda, and John Tooby. 1994. "Origins of Domain Specificity: The Evolution of Functional Organization." In *Mapping the World: Domain Specificity in Cognition and Culture*, ed. L. Hirschfield, and S. Gelman, 85-116. New York: Cambridge University Press.

Cowie, Fiona. 2008. "Innateness and Language." In *The Stanford Encyclopedia of Philosophy*, winter 2016, ed. E.N. Zalta. <http://plato.stanford.edu/archives/win2016/entries/innateness-language/>

de Almeida, Roberto G., and Lila R. Gleitman, eds. 2018. *On Concepts, Modules, and Language: Cognitive Science at its Core*. New York: Oxford University Press.

Deacon, Terrence W. 2010. "A Role for Relaxed Selection in the Evolution of the Language Capacity." *Proceedings of the National Academy of Sciences of the United States of America* 107: 9000-9006.

Dehaene, Stanislas. 2005. "Evolution of Human Cortical Circuits for Reading and Arithmetic: The 'Neuronal Recycling' Hypothesis." In *From Monkey Brain to Human Brain*, eds. Stanislas Dehaene, J.R. Duhamel, M.D. Hauser, and G. Rizzolatti, 133-157. Cambridge, MA: MIT Press.

Edelman, Gerald M., and Joseph A. Gally. 2001. "Degeneracy and Complexity in Biological Systems." *Proceedings of the National Academy of Sciences of the United States of America* 98 (24): 13763-13768.

Eliasmith, Chris. 2015. "Building a Behaving Brain." In *The Future of the Brain*, ed. Gary Marcus, and Jeremy Freeman, 125-136. Princeton: Princeton University Press.

Fedorenko, Evelina, Michael K. Behr, and Nancy Kanwisher. 2011. "Functional Specificity for High-Level Linguistic Processing in the Human Brain." *Proceedings of the National Academy of Sciences of the United States of America* 108 (39): 16428-16433.

Fedorenko, Evelina, John Duncan, and Nancy Kanwisher. 2012. "Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area." *Current Biology* 22 (21): 2059-2062.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Friston, Karl J., and Cathy J. Price. 2003. "Degeneracy and Redundancy in Cognitive Anatomy." *Trends in Cognitive Sciences* 7 (4): 151-152.

Gauthier, I., T. Curran, K.M. Curby, and D. Collins. 2003. "Perceptual Interference Supports a Non-Modular Account of Face Processing." *Nature Neuroscience* 6 (4): 428-432.

Glenberg, A.M., M. Brown, and J.R. Levin. 2007. "Enhancing Comprehension in Small Reading Groups Using a Manipulation Strategy." *Contemporary Educational Psychology* 32:389-399.

Glenberg, A.M., and M.P. Kaschak. 2002. "Grounding Language in Action." *Psychonomic Bulletin and Review* 9:558-565.

Glenberg, A.M., M. Sato, and L. Cattaneo. 2008. "Use-Induced Motor Plasticity Affects the Processing of Abstract and Concrete Language." *Current Biology* 18 (7): R290-291.

Godfrey-Smith, Peter. 2001. "Three Kinds of Adaptationism." In *Adaptationism and Optimality*, ed. Steven H. Orzack, and Elliott Sober, 335-357. Cambridge: Cambridge University Press.

Graziano, M.S.A., C.S.R. Taylor, T. Moore, and D.F. Cooke. 2002. "The Cortical Control of Movement Revisited." *Neuron* 36:349-362.

Greenfield, P.M. 1991. "Language, Tools and Brain: The Ontogeny and Phylogeny of Hierarchically Organized Sequential Behavior." *Behavioral and Brain Sciences* 14 (4): 531- 551; discussion 551-595.

Guida, Alessandro, Guillermo Campitelli, and Fernand Gobet. 2016. "Becoming an Expert: Ontogeny of Expertise as an Example of Neural Reuse." *Behavioral and Brain Sciences* 39:13-15.

Hickok, G., and David Poeppel. 2000. "Towards a functional neuroanatomy of speech perception." *Trends in Cognitive Sciences* 4 (4): 131-138.

Iriki, Atsushi, and Miki Taoka. 2012. "Triadic (ecological, neural, cognitive) niche construction: A scenario of human brain evolution extrapolating tool use and language from the control of reaching actions." *Philosophical Transactions of the Royal Society B* 367: 10-23.

Jungé, Justin A., and Daniel C. Dennett. 2010. "Multi-Use and Constraints from Original Use." *Behavioral and Brain Sciences* 33 (4): 277-278.

Kandel, E.R., and A.J. Hudspeth. 2013. "The Brain and Behavior." In *Principles of Neural Science*, ed. E.R. Kandel, J.H. Schwartz, T.M. Jessell, S.A. Siegelbaum, and A.J. Hudspeth, 5-20. New York: McGraw-Hill.

Karmiloff-Smith, Annette. 1992. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Leo, Andrea, Giulio Bernardi, Giacomo Handjaras, Daniela Bonino, Emiliano Ricciardi, and Pietro Pietrini. 2012. "Increased BOLD Variability in the Parietal Cortex and Enhanced Parieto-Occipital Connectivity During Tactile Perception in Congenitally Blind Individuals." *Neural Plasticity* 2012:1-8 doi: 10.1155/2012/720278.

Luria, A.R., L.S. Tsvetkova, and D.S. Futer. 1965. "Aphasia in a Composer (V.G. Shebalin)." *Journal of the Neurological Sciences* 2 (3): 288-292.

MacNeilage, P.F. 1998. "The Frame/Content Theory of Evolution of Speech Production." *Behavioral and Brain Sciences* 21 (4): 499-511; discussion 511-546.

Maleszka, Ryszard, Paul H. Mason, and Andrew B. Barron. 2013. "Epigenomics and the Concept of Degeneracy in Biological Systems." *Briefings in Functional Genomics* 13 (3): 191-202.

Mason, Paul H. 2010. "Degeneracy at Multiple Levels of Complexity." *Biological Theory* 5 (3): 277-288.

McGeer, Victoria. 2007. "Why Neuroscience Matters to Cognitive Neuropsychology." *Synthese* 159:347-371.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Pessoa, Luiz. 2016. "Beyond Disjoint Brain Networks: Overlapping Networks for Cognition and Emotion." *Behavioral and Brain Sciences* 39:22-24.

Peretz, Isabelle., and Max Coltheart. 2003. "Modularity of music processing." *Nature Neuroscience* 6:688-691.

Pinker, Steven, and Ray Jackendoff. 2005. "The Faculty of Language: What's Special About It?" *Cognition* 95:201-236.

Plaut, David C. 1995. "Double Dissociation Without Modularity: Evidence from Connectionist Neuropsychology." *Journal of Clinical and Experimental Psychology* 17 (2): 291-321.

Poeppel, David. 2001. "Pure Word Deafness and the Bilateral Processing of the Speech Code." *Cognitive Science* 21 (5): 679-693.

Prinz, Jesse J. 2006. "Is the Mind Really Modular?" In *Contemporary Debates in Cognitive Science*, ed. R. Stainton, 22-36. Oxford: Blackwell.

Pulvermüller, Friedmann, and Luciano Fadiga. 2010. "Active Perception: Sensorimotor Circuits as a Cortical Basis for Language." N*ature Reviews Neuroscience* 11:351-360.

Sternberg, Saul. 2011. "Modular Processes in Mind and Brain." *Cognitive Neuropsychology* 28 (3-4): 156-208.

Striem-Amit, Ella, and Amir Amedi. 2014. "Visual Cortex Extrastriate Body-Selective Area Activation in Congenitally Blind People 'Seeing' by Using Sounds." *Current Biology* 24:1-6.

Suddendorf, Thomas. 2013. *The Gap: The Science of What Separates Us from the Animals*. New York: Basic Books.

Varley, R.A., N.J.C. Klessinger, C.A.J. Romanowski, and M. Siegal. 2005. "Agrammatic But Numerate." *Proceedings of the National Academy of Sciences of the United States of America* 102:3519-3524.

Whiteacre, James M. 2010. "Degeneracy: A Link Between Evolvability, Robustness and Complexity in Biological Systems." *Theoretical Biology and Medical Modelling* 7 (6): 1-17.

Zerilli, John. 2017a. "Against the 'System' Module." *Philosophical Psychology* 30 (3): 235-250.

——2017b. "Multiple Realization and the Commensurability of Taxonomies." *Synthese* (https://doi.org/10.1007/s11229-017-1599-1).

Dissolving the Measurement Problem Is Not an Option for the Realist*

Matthias Egg, University of Bern
matthias.egg@philo.unibe.ch

December 8, 2018

This paper critically assesses the proposal that scientific realists do not need to search for a solution of the measurement problem in quantum mechanics, but should instead dismiss the problem as ill-posed. James Ladyman and Don Ross have sought to support this proposal with arguments drawn from their naturalized metaphysics and from a Bohr-inspired approach to quantum mechanics. I show that the first class of arguments is unsuccessful, because formulating the measurement problem does not depend on the metaphysical commitments which are undermined by ontic structural realism, rainforest realism, or naturalism in general. The second class of arguments is problematic due to its refusal to provide an analysis of the term "measurement". It turns out that the proposed dissolution of the measurement problem is in conflict not only with traditional forms of scientific realism but even with the rather minimal realism that Ladyman and Ross themselves defend. The paper concludes with a brief discussion of two related proposals: Healey's pragmatist approach and Bub's information-theoretic interpretation.

1 Introduction

One of the attractions of non-realist approaches to quantum mechanics (QM) is that they offer a way to dissolve the measurement problem instead of adopting a solution for it that would either have to modify the physics (such as theories with additional variables or spontaneous collapses) or drastically inflate the empirically inaccessible content of reality (such as many-worlds interpretations). Nonetheless, many metaphysicians (and some physicists) consider the abandonment of realism too high a price to pay and therefore insist that the measurement problem calls for a (realistic) solution rather than a dissolution along non-realist lines.

*To appear in *Studies in History and Philosophy of Modern Physics*

Could there be a third position that somehow combines the attractions of these two (seemingly opposing) camps? In this paper, I critically assess a recent proposal for such a position, which can be found in the writings of James Ladyman and Don Ross (2007; 2013). While describing their approach to the measurement problem as "roughly the sort of account favoured by earlier versions of the Copenhagen interpretation" (2013, 134), Ladyman and Ross do not think this amounts to instrumentalism, but rather seek to defend it as part of their brand of scientific realism, which combines ontic structural realism (OSR) with what they call "rainforest realism" (as developed in Chapters 3 and 4, respectively, of their seminal (2007) book *Every Thing Must Go*). The purpose of the present paper is to show that their Copenhagen-style dissolution of the measurement problem does not fit well with a position that presents itself as a version of scientific realism.

In order to avoid misunderstandings, let me first mention an important point of agreement between Ladyman/Ross and myself. Much of their (2013) argument is directed against the simple realism-versus-instrumentalism dichotomy (with respect to QM), which they find operative in the philosophy of David Deutsch (2011). More specifically, they argue that there is much within the formalism of QM about which one can be a realist despite not being committed to any realistic solution of the measurement problem. I largely agree with this claim, noting that one need not be an ontic structural realist to appreciate this point (cf. Cordero 2001; Saatsi forthcoming). Nor am I particularly worried about Michael Esfeld's (2013) diagnosis of OSR being only a partial realism if it does not incorporate a realist treatment of the measurement problem. [1] What I do criticize is that the dissolution of the measurement problem proposed by Ladyman and Ross undermines some specific commitments that should be part of any position deserving to be called realism (even only a partial one).

My investigation proceeds as follows. In Section 2, I will consider how the different elements of naturalized metaphysics as developed by Ladyman and Ross might seem to undermine the traditional formulation of the measurement problem, insofar as the latter depends on applying the quantum formalism to macroscopic objects. My contention will be that while naturalized metaphysics incorporates some telling arguments against traditional views of how the macroscopic realm relates to the microworld, none of them justifies viewing the measurement problem as a pseudo-problem. Section 3 will then show how Ladyman's and Ross's Bohr-inspired alternative approach to the measurement problem contradicts some widely shared realist (and, relatedly, naturalist) commitments. This by itself may not be so problematic for Ladyman and Ross, because they do not subscribe to standard realism anyway. I will therefore continue my argument in Section 4 by demonstrating that some of Ladyman's and Ross's own arguments in favor of scientific realism are in tension with their proposed dissolution of the measurement problem. The concluding Section 5 will briefly look at two related proposals and will give reasons against thinking that these have a better prospect of being compatible with realism than the original one.

[1] In fact, I endorse a version of partial realism myself (Egg 2014).

2 Quantum states of macroscopic objects

A key step in setting up the measurement problem consists in assigning a quantum state to a measurement apparatus. In the standard example (see, e.g., Myrvold 2017, Section 4), two possible final states of the apparatus are denoted by $|\text{“0”}\rangle_A$ and $|\text{“1”}\rangle_A$, corresponding to the two basis states $|0\rangle_S$ and $|1\rangle_S$ of the measured quantum system. The measurement problem then consists in making sense of the superposed final state $a|0\rangle_S|\text{“0”}\rangle_A + b|1\rangle_S|\text{“1”}\rangle_A$, which, if a and b are both nonzero, seems to contradict our experience that instrument readings for such an experiment are always *either* “0” *or* “1”.

The assumption that the quantum formalism is the appropriate tool for describing not only microscopic entities like electrons but also macroscopic objects like measurement devices is rather widespread, so it may be surprising to see Ladyman and Ross question it. Let us therefore have a close look at their reasons for doing so:

> Note that the way we set up the measurement problem relies on the idea that the state of an apparatus for measuring, say, spin in the x-direction, is a quantum state that can be represented in the usual way by a ket vector |reads ‘up’⟩. The usual rationale for treating this as a quantum state is that the apparatus is supposed to be made of a very large number of quantum particles, but nonetheless is still essentially the same kind of thing as the electron it is measuring. However, on the view of higher-order ontology sketched above (and explained in detail in [2007, Chapter 4]), there is no reason to regard the measuring device as something that exists at all from a microscopic perspective. We have also made clear our hostility to the idea that macroscopic objects are fundamentally made of microscopic ones. Hence, the application of the quantum formalism to macroscopic objects is not necessarily justified,... (Ladyman and Ross 2007, 182; quotation to be continued in Subsection 2.3 below)

This passage shows how the two elements of Ladyman's and Ross's realism motivate their reservations about extending the quantum description from the microscopic to the macroscopic domain: While OSR denies that “macroscopic objects are fundamentally made of microscopic ones”, rainforest realism proposes a view of higher-order ontology that tells against using the same kinds of description for structures at different scales. I will now discuss these two elements and then turn to a third element which shifts the focus from realism to naturalism.

2.1 Composition, OSR, and the measurement problem

What is it, exactly, that prompts Ladyman and Ross (2007, 182) to “deny that measurement devices are the mereological sums of quantum particles”?

A first reason is their dissatisfaction with traditional philosophical accounts of mereology and composition. Early on in *Every Thing Must Go*, they describe how analytic metaphysicians have been misled by their hankering after a general *a priori* notion of composition, which pays little or no attention to what we have learnt about specific

composition relations described within various sciences. In conclusion, they note: "We have no reason to believe that an abstract composition relation is anything other than an entrenched philosophical fetish" (2007, 21).

However, this kind of criticism can be dealt with rather quickly for our purpose, by noting that no abstract metaphysical composition relation is presupposed by the claim that measurement devices are composed of quantum systems. Instead, QM itself furnishes us with the rules of how systems combine to form larger systems, and it is those rules that tell us that the compound systems can be in superposed states just as the elementary systems can. Furthermore, insofar as the quantum formalism describes interactions between the parts of compound systems, the quantum mechanical composition relation also has the necessary dynamical character that Ladyman (2017, 156) finds missing in the traditional metaphysical accounts of composition.

In the same context, Ladyman mentions the renormalization group to illustrate the vast difference between the naïve metaphysical picture of composition and the intricate way in which actual condensed matter physics describes how gross matter behaves in terms of interactions between atoms, electrons, and fields. Again, the reference to scientific accounts of composition is well taken, but I do not see how it could undermine the idea that matter is composed of quantum particles. In an important sense, this latter idea provides the very motivation for applying renormalization group methods in condensed matter physics, as a tool to eliminate degrees of freedom associated with the atomic constitution of matter that are irrelevant for its macroscopic behavior. At the same time, quantum measurement devices are characterized by the fact that *some* microscopic degrees of freedom (namely the ones being measured) are *not* eliminated, but do indeed affect the device's macroscopic behavior. The renormalization group has nothing to say about these degrees of freedom and it therefore does not tell against describing the measurement device with respect to them in the way that gives rise to the measurement problem.

A further possible reason to reject the idea of macroscopic objects being composed of quantum particles is the radical difference between the way in which quantum theories describe particles and our pre-theoretic notion of particles as "little things". Recognition of this difference lies at the heart of OSR, and it is summarized by Ladyman and Ross (2013, 143) as follows:

> Quantum entanglement in particular and quantum physics in general, especially quantum field theory, show that there is *no sense at all* in which atoms or sub-atomic particles resemble little macroscopic things reduced drastically in size. In undomesticated physics, particles don't resemble any kind of entity that people had ever imagined prior to the twentieth century.

One might criticize these statements as exaggerated (for example, by referring to some experimental procedures in nanotechnology, which allow us to manipulate atoms in ways that are not so different from how we manipulate more familiar things), but that is not the point here. The question is, rather, whether any of the surprising features of quantum particles do indeed exclude the view that measuring devices are composed of them, or

in other words, whether OSR does indeed undermine the usual way of setting up the measurement problem.

The historically most important and most widely discussed issue in OSR's account of particles concerns their identity and individuality (or lack thereof; see Ladyman 2016 and references therein). This issue, however, does not seem to have much relevance for the pertinence of viewing macroscopic objects as composed of quantum systems. On the contrary, the fact that there is a metaphysical debate on the (non-)individuality of quantum particles at all precisely shows that the quantum mechanics of composite systems is to some extent insensitive to whether the components are regarded as individuals or not. It is only their cardinality that matters, and this latter is unproblematic in non-relativistic QM (I will turn to relativistic quantum theory in the next subsection). Another way to make the same point is to note that the formalism of quantum mechanics already incorporates the features that fueled the debate on identity and individuality (namely, the indistinguishability postulate in quantum statistics and the non-supervenience of entanglement relations), so one should not expect the results of that debate to undermine the quantum mechanical account of composition.

Neither is it relevant whether particles are regarded as elementary (or fundamental) in any strong sense. I fully agree with Ladyman (2016, 202) that we should not regard them in this way, but nothing in the usual formulation of the measurement problem depends on doing so. In fact, there is excellent empirical evidence for the occurrence of superposed states in unambiguously non-elementary quantum systems (see Arndt and Hornberger 2014 for a recent review), so the whole problem can be set up without any reference to elementary or fundamental particles.

Correspondingly, admitting the measurement problem as a real problem does not presuppose taking QM as the final word on the fundamental ontology of the world. Future physics may indeed present us with fundamental theories of a completely different structure, but whatever these theories look like, they will have to accommodate those basic features of QM which allow us to predict the well-confirmed interference phenomena that motivated the development of QM in the first place. Quantum superposition is one of these features, and none of the arguments discussed so far gives us any reason to suppose that it disappears when many quantum systems combine (according to the rules of QM itself) to form a macroscopic apparatus.

2.2 Rainforest realism: particles and measurement devices as real patterns

While OSR is mostly concerned with what particles are *not* (namely, fundamental and self-subsistent individuals), rainforest realism tells us what they *are*: just like other elements of higher-order ontology, particles are real patterns (Ladyman 2016, 203-204). The notion of a real pattern is adopted from Dennett (1991) and discussed in detail in Ladyman and Ross (2007, Chapter 4), but for present purposes, the following elucidation (due to Wallace 2003, 93) will suffice: "the existence of a pattern as a real thing depends on the usefulness—in particular, the explanatory power and predictive reliability—of theories which admit that pattern in their ontology".

Recognizing particles as real patterns (rather than fundamental entities) is important,

because the particle concept becomes increasingly problematic when we turn from non-relativistic quantum mechanics to relativistic quantum field theory. The appearance of particle creation and annihilation not only undermines the above-mentioned appeal to a well-defined cardinality of particles in a composite system, but it also blurs the line between particles as persisting objects and mere excitations of quantum fields. More precisely, this latter distinction now becomes dependent on the time and energy scale at which a system is considered (Ladyman 2017, 158). The merit of rainforest realism is that it takes this dependence into account and makes room for scale-relative ontological commitments.

This implies that reference to particles is unproblematic as long as the context is appropriately specified in terms of the relevant time and energy scale. Formulations of the measurement problem implicitly do this by involving only two kinds of physical systems, namely non-relativistic quantum particles and macroscopic measurement devices. Neither of them requires consideration of quantum field theoretic effects, hence the scale at which the measurement problem is formulated is not affected by the breakdown of the particle concept in quantum field theory.

Now the notion of real patterns does not only apply to particles but also to measurement devices, and this opens another way to criticize the setting up of the measurement problem. Continuing a statement already quoted above, Ladyman and Ross (2007, 182–183) write:

> In sum, then, we deny that measurement devices are the mereological sums of quantum particles. Rather, they are real patterns and their states are legitimate posits of science in so far as they enable us to keep track of the phenomena. They do not enable us to do this if we regard them as quantum states, and therefore so regarding them is not warranted.

The idea here seems to be that QM has a more limited domain of application than supposed by those who view measurement devices as composed of quantum particles: there is an autonomous science of the macroworld in the sense that QM simply doesn't apply to macroscopic measurement devices.[2]

However, some care is needed in specifying what exactly is meant by this supposed autonomy of the macroworld. Obviously, one cannot completely disregard QM when describing the states of macroscopic measurement devices. Reiterating a point made above in the context of renormalization: the very point of performing a quantum measurement is the realization that the macroscopic domain is *not* autonomous, but influenced by some microscopic degrees of freedom (the ones being measured). Ladyman and Ross are, of course, aware of this, and they propose the Born rule as a sufficient description of how the quantum domain affects the macroscopic domain. This proposal will be discussed in Section 3. For the moment, let us just note that rainforest realism (and in particular, the idea of scale-relative ontology) should not be taken to imply an unqualified autonomy of the macroscopic scale.

[2] I owe this suggestion to an anonymous referee.

2.3 Naturalism and the assignment of quantum states

The arguments discussed so far in this section were mainly based on theoretical or even metaphysical considerations. We now turn to a more empirically oriented line of argument that involves Ladyman's and Ross's commitment to naturalism. The following passage is a direct continuation of the one quoted at the beginning of this section:

> Hence, the application of the quantum formalism to macroscopic objects is not necessarily justified, especially if those objects are importantly different from microscopic objects, as indeed they are, in not being carefully isolated from the environment. From the point of view of the [Principle of Naturalistic Closure], the representation of macroscopic objects using quantum states can only be justified on the basis of its explanatory and predictive power and it has neither. In fact, QM is explanator[il]y and predictively inaccurate at this scale since it entails that there ought to be superpositions that are not in fact observed. The predictive success of QM in this context consists in the successful application of the Born rule, and that is bought at the cost of a pragmatic splitting of the world into system and apparatus. (Ladyman and Ross 2007, 182)

The suggestion that a quantum description becomes inappropriate to the extent that systems fail to be isolated has some initial plausibility, as the interference effects indicating quantum behavior are indeed only observed for systems that are sufficiently well isolated from their environment. It is, however, questionable whether this serves as a general argument against applying the quantum formalism to macroscopic, non-isolated objects. First, one might ask whether QM is not required to explain various effects in such objects, ranging from superconductivity and superfluidity to such everyday phenomena as semiconductivity or ferromagnetism. I will not dwell on these examples because they would require a rather technical analysis of whether the explanatory and predictive success of QM in each case really depends on representing macroscopic objects using quantum states or (as in the case of measurement) on the successful application of some pragmatic recipe such as the Born rule. And even if it could be shown that the former is the case, Ladyman and Ross could still reply that the measurement case differs relevantly from the other cases, due to the manifest explanatory and predictive inadequacy of applying the quantum formalism to the measurement device.

What is needed, therefore, is some justification for assigning quantum states to a sufficiently general class of macroscopic objects that includes measurement devices. Obviously, such justification cannot be directly empirical, because there is indeed a sense in which QM (without some kind of projection postulate) "entails that there ought to be superpositions that are not in fact observed".[3] There is, however, a kind of indirect

[3] This should not be taken to suggest that we could experimentally test for interference between the terms of a macroscopic superposition and thereby falsify a prediction of QM. What is meant here by the failure to "observe superpositions" is merely the fact that, for example, we never seem to observe a measuring device displaying two different results of a single measurement at once. (Thanks to Wayne Myrvold for urging me to clarify this point.)

empirical justification that should also be acceptable to the naturalist and that becomes apparent as soon as we inquire into how the lack of isolation influences the behavior of quantum systems. Our present most successful treatment of this issue is through the theory of decoherence (see Bacciagaluppi 2016 for a review). By operating entirely within the formalism of QM, this theory presupposes just what Ladyman and Ross seek to prohibit: the assignment of quantum states to macroscopic objects (the environment).[4] Such assignments are therefore not just a philosopher's fancy, but are soundly rooted in scientific theorizing.

Ladyman and Ross (2007, 176–177) discuss decoherence theory in the context of what they call "Everett-Saunders-Wallace quantum mechanics" (a version of the many-worlds interpretation). They are somewhat sympathetic to that interpretation, but do not endorse it, because they are "not yet convinced that an Everettian plurality is the most consilient way of looking at contemporary physics" (182). So if decoherence theory only played a role in the Everettian approach to QM, Ladyman and Ross could legitimately disregard it and its practice of assigning quantum states to macroscopic objects. However, decoherence theory is by no means confined to an Everettian approach; its importance extends to all the major approaches to QM (Bacciagaluppi 2016, Section 3), and some of its central results are arguably interpretation-independent (Rosaler 2016). This is the kind of scientific theory that a naturalist should take ontologically seriously, which, in this case, means to accept the legitimacy of applying the quantum formalism to macroscopic objects.

3 Leaving "measurement" unanalyzed?

Having discussed Ladyman's and Ross's view on how *not* to apply QM to the macroscopic realm, let me now turn to the more positive part of their proposal. It starts with the remark (already quoted in Subsection 2.3) that the success of QM in a measurement context consists in the successful application of the Born rule. This is indeed successful if we simply insist (as Niels Bohr famously did) that the apparatus needs to be described classically in the sense of not being in any superposed state. Ladyman and Ross (2013, 134) explicitly sympathize with Bohr's early version of the Copenhagen interpretation, which they view as distinct from later versions in virtue of its refusal to give any story about collapse of the wave function.

Without entering into the complex debate on the history of the Copenhagen interpretation, it is noteworthy that Ladyman and Ross (ibid.) identify "an abandonment not so much of realism as of naturalism itself" in the transition from Bohr to later versions of Copenhagen, which "*did* include a story about collapse, but interpreted it as a consequence of measurement". Against this assessment, I submit that the abandonment of naturalism (and realism) takes place when one endorses Ladyman's and Ross's reading of early Copenhagen, not when one switches from this version to a later one.

[4] As is well known, the fact that decoherence theory is just an application of standard QM also implies that decoherence by itself does not solve the measurement problem (Bacciagaluppi 2016, Section 2.1).

Ladyman's and Ross's argument for viewing Bohr's approach as compatible with scientific realism depends on the interpretation-independent content of standard QM already mentioned in Section 1, where I admitted that one can be a realist about large parts of QM without committing to any particular solution of the measurement problem. However, as we just saw, some of that content gets its empirical character only via the successful application of the Born rule — the content is interpretation-independent precisely because any viable interpretation of QM needs to incorporate the Born rule in some way.

Now the problem with the Born rule is that it speaks about the probabilities of measurement results, while it is notoriously unclear what counts as a "measurement". This critique is well known, and it is often put in terms of awkward questions for those who (in the spirit of later Copenhagen) tie the notion of measurement to a collapse of the wave function. So for example, John S. Bell (1990, 19) famously asked whether a single-celled living creature already qualified to play the role of "measurer" or whether it takes some better qualified system (with a Ph.D.?) to make the wave function collapse. But the basic point of criticism (as Bell makes clear in the rest of his paper) does not depend on any specific view of wave function collapse, but on using such a desperately imprecise notion as "measurement" in a basic assumption of physics. This is why realistic versions of QM (such as those associated with the names of Everett, Bohm, or GRW) seek to *derive* the Born rule by giving a physical account of what it is to be a measurement. Bohr, on the other hand, denies the need for such an account, as Ladyman and Ross (2013, 134) point out approvingly.

Admittedly, any theory has to operate with some basic notions which are not amenable to further analysis, so why not simply treat "measurement" as such a notion? This works well for situations in which we all agree whether the notion applies or not. But what about ambiguous cases, for example, a device that displays a measurement outcome which is not (even indirectly) observed[5] by anyone? Bohr repeatedly insisted that human observers play no essential role in the measurement process, but how can this be justified without an analysis of "measurement"? Neither our pre-theoretical nor our scientific usage of the word "measurement" seems to settle the question whether unobserved measurements should still count as measurements.

There are two possible lines of response to this problem, both of which are to some extent explored by Ladyman and Ross. First, one might argue that even in the absence of a physical analysis of "measurement", the notion can be made sufficiently precise in information-theoretic terms. I will briefly comment on this idea in Subsection 5.2 below. The second option is to go verificationist and to denounce any question about unobserved measurements as a pseudo-question: Such events (by definition) do not make any difference to what we observe, hence we should not suppose that there are any matters of fact concerning them. However, this is hard to square with realism, understood as a stance that refuses to limit reality to what we can observe, or worse still, to what we actually *do* observe. Ladyman and Ross (2007, 309) are quite honest about how their

[5] A useful explication of the notion of "observation" relevant for this context is given by Ladyman and Ross (2007, 307) in terms of "informational connectedness". In the following, I have this rather wide sense of "observation" in mind when I speak of "unobserved measurements" or "unobserved data".

verificationism limits the domain of what counts as real, but the conflict with realism is obscured by the somewhat far-fetched example they give in that context: Most realists will readily agree that "there are no grounds for regarding the other side of [the Big Bang] as part of reality" (ibid.). By contrast, many realists will think that something has gone deeply wrong if we are discouraged from believing that there is a fact of the matter as to how our measurement devices behave when no one watches them.

Before I turn (in the next section) to the question in how far Ladyman's and Ross's own version of realism should be bothered by this tension, I should also mention that there is something anti-naturalistic about drawing such anthropocentric limits around what counts as real. A thorough discussion of Ladyman's and Ross's (2007, Section 6.3) arguments for the compatibility of naturalism and verificationism is beyond the scope of this paper. Suffice it to say that these arguments are most plausible when the verifiability criterion is understood epistemically (as a policy on what our theorizing should or should not be concerned with) rather than metaphysically (as a criterion on what does or does not belong to reality). To the extent that the latter reading is implied, a naturalist is likely to wonder why reality should care which parts of it are accessible to our observation.

4 Unobserved measurements and objective modality

That the Bohrian approach to the measurement problem entails a conflict with some widespread realistic and naturalistic intuitions may not be a decisive reason against it, especially if one has already abandoned certain commitments of standard realism, as proponents of OSR have. I will therefore now show that the proposed dissolution of the measurement problem conflicts not only with standard realism but also with elements of scientific realism that Ladyman and Ross themselves endorse.

A first hint of this conflict appears in the role that the notion of "data" plays within rainforest realism. As we saw in Section 2, Ladyman and Ross conceive of reality in terms of real patterns, and patterns are "relations among data" (2007, 228). In their discussion of Dennett's (1991) account of real patterns, Ladyman and Ross carefully distance themselves from the kind of instrumentalism that is at least partly invited by Dennett's writing and has preoccupied many of his commentators. In the process of doing so, they acknowledge that "there are (presumably) real patterns in lifeless parts of the universe that no actual observer will ever reach" (Ladyman and Ross 2007, 203). But such realism about patterns presupposes realism about data regardless of whether they are observed or not. The same kind of realism is required for the Peirce-inspired view of reality as "the totality of non-redundant statistics" that is developed in Ladyman and Ross (2013, Section 5) and will briefly be discussed in Subsection 5.2 below. This already shows quite clearly that Ladyman and Ross cannot accept the non-realism about unobserved quantum measurements that would follow from a strict verificationism with respect to QM.

The problem comes into sharper focus when we turn to Ladyman's and Ross's (2007, Subsection 2.3.2) critique of van Fraassen's constructive empiricism. While they largely share van Fraassen's aversion to traditional metaphysics, they defend a commitment to

objective modality as a crucial element of realism against his deflationary view. One reason for this is that "theories are always modalized in the sense that they allow for a variety of different initial conditions or background assumptions rather than just the actual ones, and so describe counterfactual states of affairs" (110). A constructive empiricist might regard the claim that science gives us knowledge about non-actual states of affairs as unjustified, because all we ever experience is the actual. But this, according to Ladyman and Ross, neglects the fact that we can to some extent vary what becomes actual and still experience that our theories accurately predict what we observe. In other words, the empiricist relies on a somewhat arbitrary boundary when confining the content of our theories to a description of what actually occurs.

Insofar as this accurately describes the motivation for preferring OSR to constructive empiricism, an adherent of OSR should be equally dissatisfied with versions of QM which fail to give a non-anthropocentric account of measurement, because they involve a similar boundary between what our theories do and do not tell us. In this case, it is not the boundary between what actually occurred and what could have occurred under different initial conditions (if the Born rule is modalized in the above sense, it does give us knowledge about both of these), but the boundary between what was actually observed and what actually occurred without being observed (the Born rule being silent about the latter set of events). This second boundary is just as arbitrary as the first one, because it is largely up to us which occurrent events are observed and which ones are not.

The same point can also be made in terms of demands for explanation. In general, Ladyman and Ross share van Fraassen's s skepticism towards such demands, but here is one they explicitly accept: "That we are so often able to identify regularities in phenomena and then use them for prediction needs to be explained" (Ladyman and Ross 2007, 106). If OSR is to have any advantage over constructive empiricism, the sought-after explanation cannot simply be that there are such regularities, because that explanation would be available to the constructive empiricist as well. In order to satisfy OSR's demand, the regularities need to be invested with modal force, which enables us to answer questions about counterfactual situations. Among such questions are those about what would have happened if we had not been around to observe the phenomena in question, and an explanation would hardly be deemed satisfactory if it postulated regularities that only obtain if some observer is present. But this is precisely what the Born rule does, if it is interpreted as a modally charged law but not supplemented by a non-anthropocentric account of "measurement".

5 Conclusion and remarks about related proposals

I have argued (in Section 2) that neither the arguments for OSR or rainforest realism nor a commitment to naturalism serve to undermine the view that measurement devices are composed of quantum systems in the sense relevant for formulating the measurement problem. Furthermore, we saw that the proposal to dissolve the measurement problem along Bohrian lines conflicts not only with some commitments of standard realism (as demonstrated in Section 3) but also with the rather minimal kind of realism that

Ladyman and Ross defend against van Fraassen (Section 4).

Some will object that this is very far from establishing the general claim contained in the title of this paper. Indeed, given the wide range of positions encompassed by the term "realism", it would be hopeless to deny that some proposed dissolutions of the measurement problem are compatible with some forms of realism. Let me nevertheless add a few remarks pointing towards a generalization of the above discussion. The two proposals I'm going to discuss have been suggested to me by an anonymous referee as possible interpretations of what Ladyman and Ross are aiming at. I will therefore not only examine these proposals with respect to their compatibility with realism but also ask whether they can legitimately be viewed as performing a similar task to the one that Ladyman and Ross have set for themselves.

5.1 The pragmatist approach

The first proposal is Richard Healey's (2017) pragmatist approach to QM. On the last pages of his book, Healey discusses the relation of his proposal to scientific realism and ultimately leaves it "to the reader to decide whether to classify the resulting view of quantum theory as realist as well as pragmatist" (257). In more recent work, however, he explicitly advertises his position as a version of realism (Healey forthcoming). The key idea is that QM, although it should not be viewed as a representation or description of the world, nevertheless gives us objective guidance for belief in certain *magnitude claims*, and these are the ones that truly describe (parts of) the world. Without going into further detail, let me just mention that one reason why a realist might not be satisfied with this approach is that it involves a rather peculiar contextualism about the content of magnitude claims (see Lewis forthcoming).

Be that as it may, what emerges quite explicitly from Healey's account is its incompatibility with the metaphysical project of Ladyman and Ross. Much of the support that QM lends to OSR depends on viewing quantum states (in particular, entangled states) as representing physical structure (although maybe not in the macroscopic realm, as we saw in Section 2), which is precisely what Healey denies. This leads him to conclude that "Quantum theory does not itself posit any novel physical structures so it does not make a contribution to fundamental physics, understood in Ladyman and Ross's way" (2017, 242). In other words, if Healey's approach is a kind of realism, then it is a realism that is even more minimal than the one defended by Ladyman and Ross. I thus suspect that it entails an even starker conflict with some standard realist and naturalist commitments than the one described in Section 3, but I admit that more work is needed to assess the severity of this conflict.

5.2 The information-theoretic interpretation

The second proposal, Jeffrey Bub's (2016) information-theoretic interpretation of QM, is more interesting, because there is quite some *prima facie* evidence that it meshes very nicely with what Ladyman and Ross are doing. In addition to some approving comments on Bub's earlier work in Ladyman and Ross (2007, 184–186), they extensively rely on

information theory in their articulation and defense of rainforest realism, culminating in what they call an "information-theoretic conception of existence" (2013, 121). Hence, one might think that they would certainly join Bub in rejecting the "big" measurement problem (in the sense of Pitowsky 2006) as a pseudo-problem, which would leave us with only the "small" measurement problem of explaining how quantum probabilities get transformed into a classical probability distribution over macroscopic measurement outcomes.

I see two reasons for being suspicious of this connection. First, as Bub and Pitowsky (2010, 438) explain, rejecting the "big" measurement problem involves rejecting "two dogmas" about QM. While it is obvious that Ladyman and Ross do indeed reject the first dogma (namely, "that measurement should never be introduced as a primitive process in a fundamental mechanical theory"), this is not so clear with respect to the second dogma, "the view that the quantum state has an ontological significance analogous to the ontological significance of the classical state as the 'truthmaker' for propositions about the occurrence and non-occurrence of events, i.e., that the quantum state is a representation of physical reality". Although they might criticize talk of "truthmakers of propositions" (because it smacks of the syntactic view of theories, which they reject), we saw above that Ladyman and Ross are committed to a representational role of quantum states (as long as we do not apply them to macroscopic objects). The second reason for suspicion is that they do not seem to be concerned with anything that resembles Pitowsky's "small" measurement problem. The closest they get to discussing it is their account of decoherence mentioned in subsection 2.3 above, which, as we saw, takes place in the context of the Everett-interpretation, a view that Bub (2016, 223) recognizes as a proposal to solve the "big" measurement problem.

But even if Ladyman and Ross could be interpreted as replacing the "big" measurement problem with the "small" one, this would do nothing to render the dissolution of the measurement problem acceptable to the realist. The resolute rejection of what Bub and Pitowsky call the "first dogma" still entails all the anti-realistic consequences discussed in Sections 3 and 4. Furthermore, it is questionable whether Bub's realist rhetoric when he describes the solution of the "small" measurement problem is justified. He describes how "at the end of the measurement, the measured system Q and the measuring instrument end up in a mixed state, which can be interpreted as a classical probability distribution over definite measurement outcomes" (227). These probabilities are "measures of ignorance over several possibilities, one of which actually happens" (225). Such an account faces an objection raised by Maudlin (1995, 9–10): True, when we are in a situation of ignorance concerning a distribution of classical possibilities, we can use a mixed state to describe the corresponding probabilities. But from this it does not follow that whenever a system is in a mixed state, there is something (namely, the real state) of which we are ignorant.[6]

This is not to say that a realist account necessarily has to interpret probabilities in

[6] If the proposal is amended by brutely postulating that one of the possibilities in the (classical) probability distribution gets selected, then (as Timpson (2010) observes) the view turns into a kind of modal interpretation that adds a *value state* to the usual quantum state. But since it then fails to say anything about how this state evolves, Timpson still deems it unacceptable to the realist.

terms of ignorance. As Ladyman and Ross (2013, 146) point out, "dissatisfaction ... with irreducibly statistical fundamental physics ... has no sound motivation in either epistemology or naturalistic metaphysics". However, satisfaction with irreducible stochasticity does not entail satisfaction with a purely information-theoretic (as opposed to physical) analysis of "measurement". Adherents of spontaneous collapse models, for example, agree that reality is irreducibly stochastic, but deny that an information-theoretic understanding of the Born rule by itself provides us with a sufficiently realistic account of that stochasticity. In particular, they will ask for a physical description of the difference between information transmitting processes described by the Born rule and other such processes (in the microscopic realm) to which the Born rule apparently does not apply, as evidenced by the persistence of interference.

What this disagreement shows is that the issue of fundamental stochasticity is at least partly independent of the problems discussed in the previous sections. The information-theoretic realism captured in the slogan "the world is the totality of non-redundant statistics" (Ladyman and Ross 2013, Section 5) may offer a fascinating new framework for addressing questions of realism, but more needs to be said about the conditions under which QM provides us with data about unobserved events. For without data, there is no statistics, and hence (according to the slogan) no world. In other words, a proper solution to the measurement problem will still be required for realism, even if it is couched in information-theoretic terms.

Acknowledgments

I would like to thank James Ladyman, Andrea Oldofredi, Wayne Myrvold and two anonymous referees for helpful comments in the development of this paper. I am also indebted to audiences in Leeds, Bern, Milan (SMS 2018) and Seattle (PSA 2018) for inspiring discussions.

References

Arndt, M. and K. Hornberger (2014). Testing the limits of quantum mechanical superpositions. *Nature Physics 10*, 271–277.

Bacciagaluppi, G. (2016). The role of decoherence in quantum mechanics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2016 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2016/entries/qm-decoherence/`.

Bell, J. S. (1990). Against 'measurement'. In A. I. Miller (Ed.), *Sixty-two years of uncertainty: historical, philosophical, and physical inquiries into the foundations of quantum mechanics*, pp. 17–32. New York: Plenum Press.

Bub, J. (2016). *Bananaworld: Quantum Mechanics for Primates*. Oxford: Oxford University Press.

Bub, J. and I. Pitowsky (2010). Two dogmas about quantum mechanics. See Saunders et al. (2010), pp. 433–459.

Cordero, A. (2001). Realism and underdetermination: Some clues from the practices-up. *Philosophy of Science 68*, S301–S312.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy 88*, 27–51.

Deutsch, D. (2011). *The Beginning of Infinity*. London: Allen Lane.

Egg, M. (2014). *Scientific Realism in Particle Physics: A Causal Approach*. Boston: De Gruyter.

Esfeld, M. (2013). Ontic structural realism and the interpretation of quantum mechanics. *European Journal for Philosophy of Science 3*, 19–32.

Healey, R. (2017). *The Quantum Revolution in Philosophy*. Oxford: Oxford University Press.

Healey, R. (forthcoming). Pragmatist quantum realism. In S. French and J. Saatsi (Eds.), *Scientific Realism and the Quantum*. Oxford: Oxford University Press.

Ladyman, J. (2016). Are there individuals in physics, and if so, what are they? In A. Guay and T. Pradeu (Eds.), *Individuals Across the Sciences*, pp. 193–206. Oxford: Oxford University Press.

Ladyman, J. (2017). An apology for naturalized metaphysics. In M. H. Slater and Z. Yudell (Eds.), *Metaphysics and the Philosophy of Science: New Essays*, pp. 141–161. New York: Oxford University Press.

Ladyman, J. and D. Ross (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Ladyman, J. and D. Ross (2013). The world in the data. In D. Ross, J. Ladyman, and H. Kincaid (Eds.), *Scientific Metaphysics*, pp. 108–150. Oxford: Oxford University Press.

Lewis, P. J. (forthcoming). Quantum mechanics and its (dis)contents. In S. French and J. Saatsi (Eds.), *Scientific Realism and the Quantum*. Oxford: Oxford University Press.

Maudlin, T. (1995). Three measurement problems. *Topoi 14*, 7–15.

Myrvold, W. (2017). Philosophical issues in quantum theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2017/entries/qt-issues/`.

Pitowsky, I. (2006). Quantum mechanics as a theory of probability. In W. Demopoulos and I. Pitowsky (Eds.), *Physical Theory and its Interpretation: Essays in Honor of Jeffrey Bub*, Western Ontario Series in Philosophy of Science, pp. 213–239. Dordrecht: Springer.

Rosaler, J. (2016). Interpretation neutrality in the classical domain of quantum theory. *Studies in History and Philosophy of Modern Physics 53*, 54–72.

Saatsi, J. (forthcoming). Scientific realism meets metaphysics of quantum mechanics. In A. Cordero (Ed.), *Philosophers Think About Quantum Theory*. Springer.

Saunders, S., J. Barrett, A. Kent, and D. Wallace (Eds.) (2010). *Many Worlds? Everett, Quantum Theory, and Reality*. Oxford: Oxford University Press.

Timpson, C. (2010). Rabid dogma? comments on Bub and Pitowsky. See Saunders et al. (2010), pp. 460–466.

Wallace, D. (2003). Everett and structure. *Studies in History and Philosophy of Modern Physics 34*, 87–105.

Author: Paul L. Franco | UW-Seattle | pfranco@uw.edu

Title: "Speech Act Theory and the Multiple Aims of Science"

Abstract: I draw upon speech act theory to understand the speech acts appropriate to the multiple aims of scientific practice and the role of nonepistemic values in evaluating speech acts made relative to those aims. First, I consider work that distinguishes explanatory speech acts from descriptive speech acts within scientific practice. I then show how speech act theory provides a framework to make sense of explaining's and describing's distinct felicity conditions. Finally, I argue that if explaining aims to convey understanding to particular audiences rather than describe literally across contexts, then evaluating explanatory speech acts directed to nonscientists involves nonepistemic criteria.

Acknowledgments: Thanks to the students in my science and values class in Fall 2017, especially Abbey Willman for writing a term paper that touched on the topic here; the Philosophy of Science Reading Group at the University of Washington; the audience at PSA 2018; Kevin Elliott; and Erin Kendig for helpful discussion and/or suggestions that improved the paper.

1. Introduction

Hasok Chang "[complains] about…our [i.e., philosophers of science] habit of focusing on descriptive statements that are either products or presuppositions of scientific work, and our commitment to solving problems by investigating the logical relationships between these statements" (2014, 67–8). He argues philosophers of science should adopt "a change of focus from propositions to actions" (67). Chang suggests, "When we do pay attention to words, it would be better to remember to think of 'how to do things with words', to recall J. L. Austin's (1962) famous phrase" (68).

In this paper, I take up Chang's suggestion and argue that attending to Austin's account of the things we do with words can help us understand the multiple aims of scientific practices, the speech acts appropriate to those aims, and the roles of nonepistemic values in evaluating speech acts made relative to those aims. To do this, I first show how Austin's speech act theory provides a framework for making sense of the ways scientific representations can be used for different speech acts depending on one's aims. Second, I show that evaluating the success of these different speech acts involves looking to felicity conditions other than truth and falsity, the securing of uptake in one's audience, and, sometimes, nonepistemic values.

In §2, I consider philosophers of science working on explanation who have shifted focus from propositions to the act of explaining and relate this work to speech act theory.[1] In §3, I provide details of Austin's framework to highlight the felicity conditions of speech acts beyond truth and falsity. In §4, I consider work on the multiple aims of scientific practice, especially aims related to conveying understanding to nonscientists, and argue that evaluating speech acts appropriate to those aims involves nonepistemic values.

[1] I make no claims Chang influenced this work.

2. Things scientists do with words

2.1 Explaining

Consider some recent and not-so-recent work on scientific explanation. Andrea Woody's functional perspective motivates "a shift in focus away from explanations, as achievements, toward explaining, as a coordinated activity of communities" (2015, 80). In a similar spirit, Angela Potochnik argues that "sidelining the communicative purposes to which explanations are put is a mistake" (2016, 724). For Potochnik, explaining is a communicative act involving a speaker and audience made against a background that shapes the explanations offered. In so arguing, Potochnik deliberately recalls Peter Achinstein's claim, "Explaining is an illocutionary act," i.e., a speech act uttered by a speaker to an audience with a certain force and for a certain point (1977, 1).

These accounts share in common an emphasis on the importance of context, especially the aims of the speaker and interests of the audience in evaluating, to use Austin's terminology, the felicity conditions of explanatory speech acts. In particular, we might focus on the aims of the speaker and their audience in giving and requesting explanations, and the time and location of an explanatory speech act in deciding if the act is successful or felicitous. In focusing on the explaining act rather than the supposedly stable propositional content of an explanation, our attention is drawn to ways of evaluating that act beyond truth and falsity.

Related to this last point, Nancy Cartwright argues the functions of a scientific theory to "tell us…what is true in nature, and how we are to explain it…are entirely different functions" (1980, 159). *Ceteris paribus* laws are literally false, but still do explanatory work. One way to understand Cartwright's claim is that the speech act of describing the world truly and the speech act of explaining come apart from one another and fulfill distinct aims within scientific practice. It follows that descriptive and explanatory speech acts have different felicity conditions. If this is

right, evaluating explanatory speech acts solely in terms of truth or falsity can be inapt. For example, suppose explaining aims to increase understanding in one's audience. As Potochnik (2016) argues, what gets explained depends on a speaker's and audience's interests, and the success of an explaining act in generating understanding depends, in part, on the cognitive resources of the audience. As such, to evaluate any given act of explaining requires attending to the interests and cognitive resources of speakers and audiences and the context in which explanations are offered. This moves beyond merely focusing on the descriptive content of explanatory speech acts.

2.2 Multiple aims

A focus on acts and away from the truth or falsity of descriptive content is not unique to recent work on explanation. We see a similar shift in the aims approach to values in science, e.g., Kevin Elliott and Daniel McKaughan (2014), and Kristen Intemann (2015). The aims approach, like the mentioned work on explaining, recognizes scientific practice aims at more than describing the world and so the results of scientific practice can be evaluated from a number of perspectives related to those aims. As Elliott and McKaughan put this point, "representations can be evaluated not only on the basis of the relations that they bear to the world but also in connection with the various uses to which they are put" (2014, 3). Further, if some of those uses include things like providing timely input for policymakers or increasing public understanding of science for ethical and political reasons, there is a role for nonepistemic values in evaluating the success of those uses.

I think the general framework of Austin's speech act theory helps flesh out this picture about the multiple aims of scientific practice and their relationship to nonepistemic values in at least two ways. First, speech act theory makes sense of how one and the same sentence can be used

to perform different speech acts depending on the aims of the speaker and the context of utterance. Second, it shows that evaluating different speech acts requires more than looking at "the basis of the relations that they bear to the world" (3). Instead, to properly evaluate speech acts we have to look to the aims of the speaker and the interests of their audience, including whatever nonepistemic values are relevant to those aims and interests.

Take Austin's claim that evaluating apparently descriptive speech acts like "'France is hexagonal,'" involves questions about who is uttering the statement, in what context, and with what "intents and purposes" (1962, 142). Rather than concluding the sentence is false and leaving it at that, Austin points out the different speech acts one can use such a sentence to perform, e.g., stating or estimating. In determining the use the sentence is put to—by consulting context and inquiring after the aims of the speaker and the interests of their audience—we might realize, irrespective of the sentence's literal truth or falsity, "It is good enough for a top-ranking general, perhaps, but not for a geographer" (142). In other words, it serves the aims of the general, which, unlike the aims of the geographer, do not require a descriptively literal account of France's shape. As such, evaluating the speech act solely in terms of truth or falsity misses something important since the speaker might not be aiming to describe literally, but at something else entirely. Further, if the aims of the general are nonepistemic in character, we can evaluate the felicity of the speech act relative to how well it meets those aims.

In making these points, I think Austin is right that we can "play Old Harry with two fetishes…(1) the true/false fetish, (2) the value/fact fetish" (150). In combating these fetishes, Austin sought to free philosophers from the view "that the sole business, the sole interesting business, of any utterance…is to be true or at least false" (1970, 233). In doing so, speech act theory motivates a constructive shift from the truth or falsity of descriptive statements to

considering the multiple aims we have in performing different speech acts and the role of nonepistemic criteria in evaluating how well those speech acts meet aims not purely epistemic in character.

To expand on this picture, I turn to explicating Austin's speech act theory.

3. Speech act theory

3.1 Performatives and constatives

Austin first drew our attention to things we do with words by discussing performative utterances. Of these, Austin says, "if a person makes an utterance of this sort we should say that he is *doing* something rather than merely *saying* something" (1970, 235). Imagine a speaker utters 'I promise to return my referee report in two weeks' during the peer-review process. In promising, Austin claims the speaker does not describe an internal act she has concurrent to her utterance. Instead, in making that utterance, the speaker performs the act of promising thereby committing herself to actions related to the timely review of papers.

While promising has no special connection to truth, it still must meet certain felicity conditions to be happy. In order to successfully promise to return their referee report in two weeks, the speaker must meet the sincerity condition of forming an intention to do so and must also be able to realize their intention. There is unhappiness in, or an abuse of the speech act if the speaker promises knowing other commitments will prevent her from returning the report in two weeks. The speaker must also have the authority to make a promise; unless authorized, an editor cannot promise on behalf of a reviewer. There should also exist a convention for making promises in this context. Such conventions might allow the speaker to promise without uttering, 'I promise,' e.g.,

by accepting a request that reads, 'In agreeing to review you commit to returning your report within such-and-such a time.'

Austin first contrasts performatives with constatives, e.g., descriptive statements or assertions that aim to state something true about the world, but which do not seem to be actions. However, Austin claims describing and asserting are as much actions as promising, even if their felicity conditions are closely connected to truth and falsity. Consider an editor saying of a reviewer, 'They review quickly, and I expect they will return their review within two weeks.' In saying this, the editor commits herself to providing evidence for her description of the reviewer as quick, and perhaps justifying her expectation that the reviewer's past behavior provides good evidence for future behavior. As Robert Brandom says, "In asserting a claim one not only authorizes further assertions, but commits oneself to vindicate the original claim, showing that one is entitled to make it" (1983, 641). That is, the utterer must be in a position of authority—here in an epistemic sense—with regards to the claim and be ready to perform further speech acts if prompted. Other felicity conditions of assertions include a sincerity condition; generally, people should believe what they say. Finally, the context of an assertion shapes its felicity conditions: an editor should utter the sentence in appropriate circumstances, e.g., as a response to concerns about the speed of the reviewer. Should these conditions not be met, the speech act might be unhappy even if true.

3.2 Locution and illocution

Austin develops speech act theory to capture the similarities between performatives and constatives. Speech acts like promising and describing have three dimensions: the locutionary content, which is the conventional sense and reference of the uttered sentence; the illocutionary

force, which is the use the utterance is put to; and the perlocutionary effects, which are intended and unintended "effects upon the feelings, thoughts, or actions of the audience, or of the speaker, or of other persons" (1962, 101).

Austin's points about the illocutionary dimension of a speech act most clearly capture how a single representation can be put to different uses depending on our aims, and how different uses have different felicity conditions despite sharing locutionary content.[2] Consider the sentence, 'This product contains chemicals known to cause cancer.' The locutionary content consists in the proposition expressed by the sentence as determined by the conventional sense and reference of the words and can be common to different illocutionary acts. Someone uttering the sentence could be describing a product, issuing a warning, or explaining why they use a particular product but not another. Uttering the sentence with the force of a description, the force of a warning, or the force of an explanation will share some felicity conditions related to truth. Namely, the locutionary content should be true or approximately true to count as a good description, a good warning, or a good explanation.

However, a warning might be infelicitous in ways a description might not. For example, warnings might be issued only when a pre-determined level of significant risk at a certain level of exposure is met. In cases where such levels are not met, issuing a warning might be infelicitous. Consider also that uttering such a sentence with the force of an explanation might be called for only if, e.g., someone is prompted to justify their choice of a product that does not contain cancer-causing chemicals over a more easily available and cheaper product that does. In these last two

[2] The inductive risk argument in science and values focuses on perlocutionary effects. See Heather Douglas (2009) and Franco (2017).

cases, nonepistemic criteria related to risk, cost-effectiveness, and so on can be used to evaluate the happiness of warnings or explanations.

Austin thinks attending to these points combats a form of abstraction that distorts our thinking about the felicity conditions of speech acts. When examining descriptive statements, Austin thinks "we abstract from the illocutionary…aspects of the speech act, and we concentrate on the locutionary" (1962, 144–5). Such an approach focuses on "the ideal of what would be right to say in all circumstances, for any purpose, to any audience, &c." (145). But in doing so, "we use an over-simplified notion of correspondence with the facts—over-simplified because essentially it brings in the illocutionary aspect" (145). Questions concerning correspondence with the facts brings in the illocutionary aspect since truth or falsity does not attach to sentences or locutionary content. Instead, truth or falsity is related to particular things speakers do with words. Descriptions might be true or false, but, strictly speaking, not warnings or explanations. In order to know, then, if evaluating a speech act along the true-false dimension is apt, we need to know its illocutionary force. But to know the illocutionary force requires we attend to context, including the aims of both speaker and audience, time and place of utterance, and conventions governing the specific speech situation. In this way, Austin argues context and aims are central to determining the illocutionary force of a speech act, and hence to evaluating its felicity.

4. Aims approaches and speech act theory

4.1 Explaining and understanding

Scientific practice might seem to deal in paradigmatically constative speech acts, e.g., descriptions. Such speech acts are, to varying degrees, evaluable along dimensions of truth or falsity in ways we might question speech act theory's relevance to philosophy of science. Maybe scientific

practice just is a case in which abstracting away from illocutionary force to focus on locutionary content is appropriate. For example, Austin says "perhaps with mathematical formulas in physics books…we approximate in real life to finding" speech acts where focusing solely on the locutionary content is not pernicious (1962, 145). If scientific practice aims at timeless, true descriptions holding across all contexts independent of the aims and interests of speakers and audiences necessary to evaluating the felicity of speech acts, then perhaps speech act theory is irrelevant to philosophy of science.

Yet, as Austin points out, "When a constative is confronted with facts, we in fact appraise it in ways involving the employment of a vast array of terms which overlap with those that we use in the appraisal of performatives. In real life, as opposed to the simple situations envisaged in logical theory, one cannot always answer in a simple manner whether it is true or false" (141–2). Consider again 'France is hexagonal.' Austin asks, "How can one answer…whether it is true or false that France is hexagonal? It is just rough, and that is the right and final answer to the question of the relation of 'France is hexagonal' to France. It is a rough description; it is not a true or false one" (142). Though rough, it is still open to evaluation. We can ask if it accords with conventions governing estimations for the particular purpose it is put to and if this particular estimation serves the purposes and interests of the speaker and their audience. 'France is hexagonal' can count as felicitous even if rough and not literally true because it might aim at something other than truth.

McKaughan makes a related point about scientific speech acts. He argues certain speech acts central to scientific practice like "conjecturing, hypothesizing, guessing and the like often play a role in scientific discourse that serves neither to assert that an hypothesis is true nor to express such a belief" (2012, 89). For example, following Woody, when examining particular acts or patterns of explaining used in scientific practice we might focus not on the locutionary content,

but on the ways "explanatory discourse…functions to sculpt and subsequently perpetuate communal norms of intelligibility" (2015, 81). In focusing on this aspect of explanatory speech acts, we might find, for example, that "the ideal gas law's role in practice is not essentially descriptive, but rather prescriptive; by providing selective attention to, and simplified treatment of, certain gas properties (and their relations) and ignoring other aspects of actual gas phenomena, the ideal gas law effectively instructs chemists in how to think about gases as they are characterized within chemistry" (82). On Woody's view, the ideal gas law, in practice, does not have the force of a descriptive speech act, but lays down a rule guiding the investigation of gases.[3] The success of explanatory speech acts from this perspective has less to do with describing actual gases, and more to do with the way they facilitate, say, the education of new scientists or increase understanding of related phenomena, e.g., "by laying foundation for the concept of 'temperature'" beyond "the subjective, inherently comparative quality of human perception" (82). Depending on one's aims, an explanatory act that fails to increase understanding of related phenomena might be infelicitous even if the locutionary content confronts the facts in approximately true ways.

In a related vein, Potochnik claims "that what best facilitates understanding is not determined solely by the relationship between a representation and the world" (2015, 74). Suppose a scientist's aim is to increase understanding of some phenomena rather than to describe it in all its complexity. In this case, a particular explanatory speech act making use of the ideal gas law is not defective because it fails to describe all causal factors at play in the behavior of actual gases. An explanatory speech act making use of an idealization might successfully fulfill the aims of a scientist insofar as it "secure[s] computational tractability" or isolates "all but the most significant causal influences on a phenomenon" (71). In eschewing descriptive complexity in favor of other

[3] Austin (1962, 143) entertains a similar point about laws.

goals, we increase our understanding by facilitating "successful mastery, in some sense, of the target of understanding" or "by revealing patterns and enabling insights that would otherwise be inaccessible" (72).

Moreover, Potochnik argues, "Because understanding is a cognitive state, its achievement depends in part on the characteristics of those who seek to understand," including both the speaker and the audience (2015, 74). In evaluating an act of explaining, then, we should look at how the speaker's aims shape the focus of their explanation and also how the explanation increases an audience's understanding, where this involves considering their interests in seeking an explanation. An explanation irrelevant to the audience's interests or that fails to increase their understanding or guide their thinking about related phenomena, but that nonetheless has approximately true locutionary content might count as infelicitous.

4.2 Values

On the views of explaining canvassed, the aims of generating literally true descriptions of the world come apart from, say, explaining and understanding the most important causal factors at play for a given phenomenon. Now, as the aims approach to the role for nonepistemic values in scientific practice emphasizes, explaining and describing to fellow scientists do not exhaust the goals of scientific practice. The aims approach focuses on the ways "scientific decision-making, including methodological choices, selection of data, and choice of theories or models, are...a function of the aims that constitute the research context" (Intemann 2015, 218). Given that the research context includes social, political, and moral considerations, the aims of science are often nonepistemic in character.

Consider, for example, the American Geophysical Union's position statement on human-induced climate change. At the end of their statement, they claim, "The community of scientists has responsibilities to improve overall understanding of climate change and its impacts. Improvements will come from pursuing the research needed to understand climate change, working with stakeholders to identify relevant information, and conveying understanding clearly and accurately, both to decision makers and to the general public" (American Geophysical Union 2013). Here, I focus on the claim that scientists have responsibilities to improve the understanding of policymakers and the general public. Drawing upon the aforementioned work on explaining, I consider how this aim and the values of policymakers and the public shape the nonepistemic felicity conditions of explanatory speech acts directed at them.

Notice that the position statement distinguishes the research necessary to understand climate change from conveying that understanding to policymakers and the general public. The sense in which these activities come apart and have different felicity conditions can be made sense of, in part, by focusing on the audience to whom scientists are speaking. For Potochnik (2016), understanding is a cognitive state that depends on the abilities and interests of those explaining and those to whom explanations are directed. In communicating to specific audiences of policymakers and specific audiences composed of members of the general public, scientists should consider the interests of the audience in asking for an explanation as well as their level of knowledge regarding the phenomenon in question, in this case, climate change.[4] In so doing, scientists might find a description that describes climate change in all its complexity might not serve these aims well. Instead, scientists might aim for an explanation that, though omitting

[4] Assuming a specific audience is identifiable. See Stephen John (2015) on the difficulties of carrying out similar suggestions when no single, specific audience is identifiable.

descriptive complexity, draws upon models that include causal factors related to their audiences' interests in understanding climate change, some of which will be nonepistemic character, e.g., mitigating risks from extreme weather events. Furthermore, a scientist's speech acts should be cognitively accessible for the nonscientists in their audience, perhaps in such a way that it guides their thinking more generally about climate change and its impact on things they value.[5]

On this point, the American Geophysical Union's position statement maintains scientists ought to enlist the help of stakeholders in identifying potentially relevant information to their research. In developing the aims approach, Intemann emphasizes a similar point. She says of climate science, "[T]he aim is not only to produce accurate beliefs about the atmosphere, but to do so in a way that allows us to generate useful predictions for protecting a variety of social, economic and environmental goods that we care about" (2015, 219). In the view of the American Geophysical Union, in order to do this well, scientists ought to consult with relevant stakeholders and policymakers regarding their values. For example, if stakeholders and policymakers communicate worries about extreme weather events and ask about "how to adapt to 'worst case scenarios,' then models able to capture extreme weather events should be preferred" to models that "anticipate slow gradual changes" (Intemann 2015, 220). Notice that in making such a decision, the grounds for choosing models able to represent aspects of climate change relevant to stakeholders' interests are nonepistemic rather than epistemic, e.g., generating predictions useful for protecting goods stakeholders care about. Insofar as the explanations generated do not meet these goals because they are unrelated to stakeholders' interests, the attendant speech acts might be infelicitous even if they describe some related phenomenon more or less accurately.

[5] This suggestion could be extended to other forms of communication, e.g., visual representations like infographics.

Both points about pitching cognitively accessible explanations and choosing models for representing climate change phenomena in ways sensitive to stakeholders' values and interests illustrate Austin's emphasis on the importance of uptake to successfully performing a speech act. Austin claims, "Unless a certain effect is achieved, the illocutionary act will not have been happily, successfully performed….I cannot be said to have warned an audience unless it hears what I say and takes what I say in a certain sense….Generally the effect amounts to bringing about the understanding of the meaning and force of the locution" (1962, 116). In aiming to convey understanding through explaining relevant aspects of climate change to policymakers and the public, a speaker should consider the interests, background knowledge, and cognitive resources of their audience. Insofar as scientists fail to do so in explaining to nonscientists, they will not secure uptake in the sense of generating understanding in their audience, even if the locutionary content of their speech act approximates truth.

Of course, a scientist's explaining something to their audience will also be infelicitous if based on inaccurate information or if it extrapolates from what is known to their audience's interests in unjustified ways. However, if scientists aim to increase public understanding, they should not stick solely to descriptively complex claims, but aim at making explanatory speech acts relevant to their audience's interests in cognitively accessible ways. Elliott, for example, emphasizes the importance of securing uptake in discussing how scientists should communicate uncertainty: "It does little good to expect scientists to provide unbiased information to the public if their pronouncements are completely misinterpreted or misused by those who receive them" (2017, 89). Thus, if scientists are to meet responsibilities the American Geophysical Union claims they have with regard to conveying understanding about climate change, those scientists should communicate using explanatory speech acts best able to secure uptake in the general public and

policymakers. This involves considering the epistemic and nonepistemic interests and cognitive resources of their audience in ways that shape the felicity conditions of the speech acts beyond truth and falsity.

5. Conclusion

Speech act theory can tie together threads in recent work on explaining and the aims approach to values in science that share in common a shift in focus from descriptive propositions to other things scientists do with words. Explaining is at least one of the things scientists do with words that aims at something other than describing the world literally. When we look at, say, the aims of scientists in explaining some phenomena to nonscientists through the lens of speech act theory, our attention is drawn to ways explanatory speech acts can be happy or unhappy beyond describing truly or falsely. For example, successfully securing uptake in the general public or policymakers in ways that increases their understanding of phenomena relevant to their nonepistemic interests requires attention to the cognitive resources and values of audiences, as well as the contexts in which explanations are requested. These all shape the felicity conditions of speech acts directed to the general public or policymakers. Future work within this framework may aim to articulate in greater detail the felicity conditions of speech acts made relative to the multiple aims of scientific practice with an eye towards their connection to the nonepistemic values of speakers and audiences.

References

Achinstein, Peter. 1977. "What is an Explanation?" *American Philosophical Quarterly* 14(1):1–15.

American Geophysical Union. 2013. "Human-Induced Climate Change Requires Urgent Action." https://sciencepolicy.agu.org/files/2013/07/AGU-Climate-Change-Position-Statement_August-2013.pdf

Austin, J.L. 1962. *How to Do Things With Words*. Ed. J.O. Urmson. Oxford: Oxford University Press.

----. 1970. "Performative Utterances." *Philosophical Papers*, 2nd edition. Eds. J.O. Urmson and G.J. Warnock. Oxford: Oxford University Press: 233–252.

Brandom, Robert. 1983. "Asserting." *Nous* 17(4):637–650.

Cartwright, Nancy. 1980. "The Truth Doesn't Explain Much." *American Philosophical Quarterly* 17(2):159–163.

Chang, Hasok. 2014. "Epistemic Activities and Systems of Practice: Units of Analysis in Philosophy of Science After the Practice Turn." *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science*, eds. Léna Soler, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost. New York: Routledge: 67–79.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, Kevin. 2017. *A Tapestry of Values*. New York: Oxford University Press.

Elliott, Kevin C. and Daniel J. McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81(1):1–21

Franco, Paul L. 2017. "Assertion, Nonepistemic Values, and Scientific Practice." *Philosophy of Science* 84(1):160–180.

Intemann, Kristen. "Distinguishing Between Legitimate and Illegitimate Values in Climate Modeling." *European Journal of the Philosophy of Science* 5:217–232.

John, Stephen. 2015. "Inductive Risk and the Contexts of Communication." *Synthese* 192:79–96.

McKaughan, Daniel J. 2012. "Speech acts, attitudes, and scientific practice: Can Searle handle 'Assuming for the sake of Hypothesis'?" *Pragmatics and Cognition* 20:1:88–106.

Potochnik, Angela. 2015. "The Diverse Aims of Science." *Studies in History and Philosophy of Science Part A* 53:71–80

----. 2016. "Scientific Explanation: Putting Communication First." *Philosophy of Science*, 83:721–732.

Woody, Andrea. 2015. "Re-orienting discussions of scientific explanation: A functional perspective." *Studies in History and Philosophy of Science Part A* 52:79–87.

Garson, J. (forthcoming). There are no ahistorical theories of function. *Philosophy of Science*

Title: There Are No Ahistorical Theories of Function

Author: Justin Garson

Abstract: Theories of function are conventionally divided up into historical and ahistorical ones. Proponents of ahistorical theories often cite the *ahistoricity* of their accounts as a major virtue. Here, I argue that none of the mainstream "ahistorical" accounts are actually ahistorical. All of them refer, implicitly or explicitly, to history. In Boorse's goal-contribution account, history is latent in the idea of statistical-typicality. In the propensity theory, history is implicit in the idea of a species' natural habitat. In the causal role theory, history is required for making sense of dysfunction. I elaborate some consequences for the functions debate.

Acknowledgments: I'm grateful to audience members at PSA 2018, where I presented this material. I also thank Daniel Dennett and Paul Griffiths for useful feedback.

1. Introduction

Theories of function are conventionally divided up into two main categories, historical and ahistorical (or backwards-looking and forwards-looking). The selected effects theory (Neander 1983, 1991; Millikan 1984) is an example of a *historical* theory, but there are other historical theories, including some versions of the organizational theory (McLaughlin 2001). *Ahistorical* theories include Boorse's goal-contribution account (1976; 1977; 2002), the propensity theory (Bigelow and Pargetter 1987), and the causal role theory (Cummins 1975; Craver 2001; Hardcastle 2002). In the 1970s and 1980s, it was common to see these two sorts of theories as competing with each other, though more recently, philosophers of biology have generally adopted a pluralistic stance, and see them as capturing different aspects of ordinary biological usage (Garson 2018). Still, the validity of the basic distinction has never been seriously challenged.

Many proponents of ahistorical theories have argued that we should accept their theories precisely *on account of* their being ahistorical. In other words, their alleged ahistoricity is often touted as a significant selling point of their theories, and a strong reason to prefer them over historical ones. There are two arguments along these lines. The first argument appeals to bald intuition, and says it's just obvious that functions don't always need history. One fanciful variant of this argument appeals to science fiction cases, like swamp creatures, instant lions, and randomly-generated worlds (e.g, Boorse 1976, 74; Bigelow and Pargetter 1987, 188). But one doesn't have to go as far as science fiction to find plausible cases of ahistorical functions in biology. Many philosophers have a strong intuition that, the very first time a new biological trait emerges and begins to benefit the organism, it has a *function* even if it was never selected for (e.g., Boorse 2002, 66; Bigelow and Pargetter 1987, 195; Walsh and Ariew 1996, 498). The second argument, which is closely related, appeals to ordinary biological usage instead of intuition. It says that historical theories run against the way biologists ordinarily think and talk about functions. At least sometimes, when biologists attribute functions to traits, they neither *cite* nor *refer to* nor *think about* history or evolution (e.g., Godfrey-Smith 1993, 200; Amundson and Lauder 1994, 451; Walsh 1996, 558; Boorse 2002, 73). Hence, ahistorical theories capture important strands of real biology.

In light of the above, my thesis might come as a bit of a shock. I claim that *there are no ahistorical theories of function* – or, put more precisely, the mainstream versions of the allegedly ahistorical theories on the market aren't actually ahistorical. If we poke and prod at those theories a bit, a historical element falls out, like contraband stashed away in a suitcase. In Boorse's version of the goal-contribution account, history is explicitly embedded in his notion of a *statistically-typical* contribution to fitness. In the propensity account, history is embedded, a little less explicitly, in the idea of a species' *natural habitat*. Finally, the only way the causal-role theorist can hope to make sense of dysfunction is to appeal to history.

Before I move on, there is one big qualification I must get out of the way. One could *invent* a purely ahistorical theory of function. One could assert, for example, that *all* of a trait's effects are its functions. In fact, the biologists Bock and von Wahlert (1965, 274)

proposed a theory of function very much along these lines. This theory (pan-functionalism?) would be ahistorical, to be sure, since even if the world were created two seconds ago in pretty much its present form, things would still have effects, and so they'd still have functions. In fact, sometimes scientists actually *do* use the word "function" synonymously with "effect." They say things like, "climate change is a *function* of deforestation," or "poor academic performance is a *function* of malnutrition." But this isn't the ordinary biological use, which the theories I cite above are trying to capture. I'll come back to this point in the conclusion.

So, I need to amend my thesis slightly. Instead of saying that there are no ahistorical theories of function, I want to say that any theory of function that satisfies two very minimal, very traditional, and largely uncontroversial, adequacy conditions, is *also* a historical theory. First, the theory should capture some distinction between functions and accidents (the function of the nose is to help us breathe but not hold up glasses). Second, the theory should capture the possibility of malfunctioning or dysfunction. If my heart seizes up due to cardiac arrest, it's failing to perform its function or it's dysfunctional. All of the theorists I engage with in this paper purport to satisfy these two adequacy criteria, or something in their vicinity, so I'm not begging any questions by insisting on these conditions.

Here's the plan for the rest of the paper. The next three sections will examine Boorse's goal-contribution theory, the propensity theory, and the causal-role theory, in turn. In the conclusion, I'll draw out the big consequences for thinking about functions.

2. Boorse's Goal-Contribution Account

Boorse's view (1976; 1977; 2002), at the most general level, is a goal-contribution account. It holds that a trait's function is just its contribution to a goal. Here, I'll focus on the subclass of functions he calls *physiological* functions. For Boorse, the *physiological* function of a trait is its species-typical contribution to the survival and reproductive prospects of an organism (1977, 555; 2002, 72). (To be more precise, Boorse carves up species into subgroups based on age and sex; the function of a trait is its typical contribution to fitness within the members of that subgroup.) Though he doesn't define a corresponding notion of *dysfunction*, he defines a closely related notion of *disease*: a disease is simply a state that "reduces one or more functional abilities below typical efficiency (1977, 555)."

Neander (1991, 182) raised a now-famous objection against Boorse; she pointed out that Boorse's view, as it stands, can't make sense of pandemic dysfunction: "dysfunction can become widespread within a population…A statistical definition of biological norms implies that when a trait standardly fails to perform its function, its function ceases to be its function; so that if enough of us are stricken with disease (roughly, are dysfunctional) we cease to be diseased, which is nonsense." Pandemic dysfunctions, moreover, don't just occupy the realm of science fiction, as in P. D. James' *The Children of Men*. UV radiation poisoning in anurans is a good example of pandemic dysfunction. Sadly, climate change might create many more pandemic dysfunctions very soon. A good theory

of function shouldn't close off the possibility that all, or most, tokens of a certain trait in a certain species are dysfunctional (or as Boorse prefers, "diseased").

Intriguingly, Boorse doesn't deny the possibility of pandemic disease. Instead, he says that in order to make sense of pandemic disease, one has to appreciate function's *historical depth*. Specifically, when we consider what's "statistically typical" for a trait, we cannot just look at what is typical right now. We have to examine the trait's behavior over a slice of time that includes the present moment and reaches far back into the past: "*Obviously*, some of the species' history must be included in what is species-typical (2002, 99; my emphasis)." He tells us that this time-slice should be longer than "a lifetime or two," and might include "millennia."

This is an extraordinary admission, given that much of Boorse's core argument for his view was propped up on the claim that both biology and intuition need purely ahistorical functions, uncluttered by history. His admission implies that his two key arguments for the view don't work. First, by his own lights, it's not the case that biologists don't refer to history; implicitly, when they talk about what's statistically-typical, they *are* talking about history. Second, regardless of whether or not intuition supports ahistorical functions, Boorse's theory doesn't. It's just not true, on Boorse's account, that if lions popped into being from an unparalleled saltation, their distinctive parts and processes would have functions. They wouldn't, since they don't have the right history (or to be more precise, they have no history at all).

3. The Propensity Theory

Bigelow and Pargetter (1987) also developed an influential "ahistorical" theory of function, the propensity theory. They reject the selected effects theory (and etiological accounts more generally) because the selected effects theory gets the *modality* of functions wrong. In other words, the statement, "functions are selected effects," if true, is contingently true; it might be true on the actual world, but there are possible worlds at which it's false. To illustrate the point, they ask us to consider a world that is pretty much the same as ours except that it randomly popped into being five minutes ago. On that world, they claim, there would still be functions, just no selected effects (188): "we have the intuition that the concept of biological function…[is] not thus contingent upon the acceptance of the theory of evolution by natural selection." This consideration prompts the need for an ahistorical theory.

For Bigelow and Pargetter, functions are propensities, or probabilistic dispositions. We might quibble over what exactly dispositions are, but any good definition will cite three parts: structure, environment, and behavior. Consider the solubility of salt. There is a *structure*, namely, the polar molecular structure composed of sodium and chloride; there is an *environment*, namely, water; there is a *behavior*, namely, dissolving. When we say that salt is disposed to dissolve in water, we're saying that, if you were to take something with this structure, and put it in this environment, it would perform this behavior, all things equal.

Functions, too, are dispositions. Consider "the function of the heart is to circulate blood." For this statement to be true, there must be a structure (the heart, embedded the right way in the circulatory system), an environment (which they call the creature's *natural habitat*), and a behavior (conferring a fitness boost on the organism). If one were to put the structure in its natural habitat, it would increase the fitness of the organism (relative, I suppose, to creatures without hearts). The crucial distinction between their view and Boorse's is that in their view, a trait's function doesn't depend on actual frequencies of performance. A trait needn't have an actual track record of boosting fitness to have a function; a mere propensity will do.

This raises the thorny question of what a creature's *natural habitat* is. For they're clear that a creature's natural habitat isn't just any environment the creature happens to find itself in. Curiously, they refuse to define this crucial notion; instead, they brush it off as vague, but unproblematically so: "there may be room for disagreement about what counts as a creature's 'natural habitat;' but this sort of variable parameter is a common feature of many useful scientific concepts" (192). But one could at least form the suspicion that if one analyzed this unproblematically vague notion, one would find some reference to history tucked away inside of it.

This suspicion is confirmed in the very next paragraph of their paper. There, they tell us that, if a creature's environment were to change very suddenly, then "natural habitat" will still refer to the *old* environment, and not the *new* one (ibid). There's a time lag built into the very idea of a natural habitat. So, for example, if climate change melts enough Arctic ice, then, at least for a time, the polar bear's natural habitat (and by extension, the natural habitat of the trait itself, namely, their thick, water-repellant fur) is the icy habitat of yore and not the contemporary, denuded one. They take that as given, and I agree.

But why would this be? What *makes it the case* that, in cases of rapid habitat change, "natural habitat," at least for a time, refers to the old environment and not the new one? What makes it true, I suspect, is that the idea of a natural habitat is an intrinsically historical notion. It's something like *the environment within which the species recently survived and thrived*. And if that's not what a natural habitat is, I would like to know what it is *such that*, if a creature's actual habitat shifts suddenly, the natural habitat, for a little while, is still the old one. Just because a concept is vague around the edges, that doesn't excuse one from the obligation to give some sort of analysis. Perhaps one could revise the theory and drop all reference to "natural habitat," as suggested by Griffiths (2009, 27), but that remains to be worked out in a rigorous way. Moreover, it's not clear whether such a theory, when rigorously developed, would hang together with the two adequacy criteria.

Hence, I conclude that, contrary to rumor, the propensity theory is not an ahistorical theory, or not demonstrably so. But if that's right, proponents of the propensity theory lose one of the main virtues of the view, which is to get the modality of functions right. To be fair, there's still a sense in which their view *is* ahistorical. What they can do, that the selected effects theorist can't, is to attribute functions to novel traits – so long as that

novel trait belongs to the members of a species that has been around long enough to have a natural habitat. Suppose a gene mutation confers a benefit on an organism, say, pesticide resistance in a flour beetle. I suppose they can say that, at the very moment at which it first confers that benefit, the gene mutation has a function, namely, to make the beetle withstand a certain pesticide. This result, they claim, is "intuitively comfortable" (195). But they can say that only because flour beetles themselves have a history, and so we can talk meaningfully about their natural habitats. Moreover, I think they'll still have a rough time explaining dysfunction (Neander 1991, 183), for reasons I'll point to in the next section. Finally, I think there are good theory-neutral reasons for saying that beneficial traits, on their very first appearance, don't have functions, but rather, whatever benefit they bring is a lucky accident. But I won't argue for that here (see Garson 2019, Chapter 2).

4. The Causal Role Theory

What about the causal role theory of function? This appears to be a purely ahistorical view. The causal role theory says, roughly, that the function of a *component* of a system consists in its contribution, in tandem with the other components, to a system-level capacity of interest (Cummins 1975; Craver 2001; Hardcastle 2002). Craver (2001) helpfully elaborates this view by specifying that the part in question must be a component of a *mechanism*. All of the basic ingredients of this theory, it seems, are ahistorical: capacities, components, organization, hierarchy, interests. Even if the world were created five minutes ago, in pretty much its present form, things would still have causal role functions.

The problem enters when we think about dysfunction. Cummins (1975, 758) insisted that functions are dispositions, or capacities: "…to attribute a function to something is, in part, to attribute a disposition to it." The function of a trait *token*, then, consists in its capacity to contribute to a system-level effect. But what if the token in question, through defect or disease, loses the capacity, and so can't contribute to the system-level effect? Then, by Cummins' analysis, it doesn't have the relevant function – so it can't be dysfunctional either.

Causal role theorists have, by and large, been silent about how to make sense of dysfunctions from this perspective. Almost everything they've had to say on that score, however, is consistent with the following theme: a trait *token* is dysfunctional when it can't do what other trait tokens generally, or typically, do to contribute to the system-level effect of interest. Consider Godfrey-Smith (1993, 200): "Although it is not always appreciated, the distinction between function and *mal*function can be made within Cummins' framework…If a token of a component of a system is not able to do whatever it is that other tokens do, that plays a distinguished role in the explanation of the capacities of the broader system, then that token component is malfunctional." Craver (2001, 72), offers the same general line: "…the ascription of a function to a malformed or broken part is derivative upon a description of how that *type* of part (X) fits into a *type* of higher-level mechanism (S). The malformed and broken part can be identified as an X by

the typical properties and activities of Xs....” This is, at root, to rely on a statistical norm for making sense of dysfunction.

This account of dysfunction, like Boorse’s, stumbles when it encounters the problem of pandemic dysfunction. For the modification suggested above implies that, if everyone’s heart seized up at once, nobody’s heart would have a function anymore, so nobody’s heart would be dysfunctional. The best way to solve this problem, and perhaps the only way, is the way Boorse took, namely, to say that the function of a trait is its typical contribution to some system effect, where what’s typical is assessed over a chunk of time that stretches back into the past, for at least “a lifetime or two,” and perhaps “millennia.” But if causal role theorists take that line, they’d have a historical theory.

Craver (2001) and Hardcastle (2002) suggest, all too fleetingly, a different way of thinking about dysfunction, one that depends not on statistics, but on our values, that is, the values and goals of people who make function attributions. Craver (2001, 72) suggests that traits are dysfunctional when they cannot do what people *want* them to do: “the mechanistic role of the broken part only appears against the fixed backdrop of shared assumptions about a type of mechanism within which parts of this type generally (or preferably) make important contributions.” The parenthetical remark alludes to a substantially new doctrine, one that demands our full concentration. It suggests that dysfunction is a mirror of human preferences and goals, of our wishing and wanting. If my heart seizes up, it’s dysfunctional, since it’s not doing *what I want it to do*.

Hardcastle (2002) makes remarks along similar lines. She first says that the function of a trait – what it’s “supposed to do,” as she puts it – depends on the goals of the scientific discipline that makes the investigation: “The teleological goal for some trait…depends upon the discipline generating the inquiry” (153). The palmomental reflex causes a chin twitch when you stroke an infant’s palm; it’s just an accident of cortical wiring with no deep evolutionary rationale. Still, she says, it has the *function* of indicating the state of brain development in infants, because that’s how biomedical researchers use it. She then says that something is malfunctional just when it cannot do what it’s supposed to do (152). The palmomental reflex is malfunctional when it can’t indicate the state of brain development. Simply put, dysfunction happens when a trait can’t do what we want.

But dysfunctions can’t be reduced to mere preferences in any straightforward way; this is a point that’s been taken in the literature for decades (e.g., Boorse 1977, 544; Wakefield 1992, 372), for reasons that scarcely need to be rehearsed. I’d prefer not to need sleep and water; I’d prefer if nobody had to go through the pain of childbirth or teething, either. But none of those things are dysfunctions. For that matter, I’d prefer if my hands were equipped with retractable adamantium claws. The fact that my hands can’t do what I want them to do doesn’t make them dysfunctional. If one really wanted to run with this value-centered line about dysfunction, one would *at least* have to add that, in order for a trait to be dysfunctional, it’s not enough that it doesn’t do what I prefer, but I must also have a *reasonable expectation* that it *should* act in the way that I prefer. But what could possibly ground a *reasonable expectation* that my hand (say) work in a certain way? Only this: that hands usually *do* work in the preferred way. But then we’re back to statistical

norms, and long historical slices of time. This value analysis of dysfunction isn't a contender to a statistical analysis; instead, the former presupposes the latter.

I've walked through three allegedly ahistorical theories of function, and shown that none of them are purely ahistorical; they're *infected* with history. The conclusion will say what we should do next.

5. Conclusion

There are no ahistorical theories of function, at least among the mainstream theories that are put forward as ahistorical. The first, Boorse's goal-contribution theory, explicitly refers to what's statistically typical for a trait, where what's typical is assessed over a long historical period of time. The second, the propensity theory, refers to the creature's natural habitat, which is implicitly historical. And the third, the causal role theory, can't hope to make sense of dysfunction (or so I argue) without appealing to a statistical norm, and thereby (following Boorse) to history. None of these theories will give functions to the parts of swamp creatures, instant lions, or anything on worlds that are similar to ours except for being randomly generated five minutes ago. The propensity theory, at least, can give functions to novel traits as soon as those traits begin benefiting their bearers, as long as the population in which the traits emerge has been around for long enough to have something like a natural habitat. But even that theory will probably encounter problems when it comes to making sense of dysfunction, though I haven't pushed that line in any detail here.

If my thesis is correct – that there are no ahistorical theories of function – three consequences immediately follow. First, we need to jettison this whole way of dividing up theories of function. The distinction between etiological and non-etiological theories serves us much better. An *etiological* theory holds that function ascriptions either are, or purport to be, causal explanations for the existence of traits. Non-etiological theories hold that function ascriptions are not, and they don't purport to be, causal explanations for traits. But the crucial point is that being etiological and being non-etiological are just *two different ways of being historical*.

Second, given that there are no ahistorical views, the two main arguments that have repeatedly been put forward for those theories – the argument from intuition and the argument from ordinary biological usage – don't actually work. If we took those arguments seriously, they'd count as evidence *against* these allegedly ahistorical theories. That doesn't mean those theories are wrong. It does mean, however, that we need to rethink, from the ground up, the motivation for accepting those theories.

A third consequence is that one popular way of thinking about function pluralism must fail. This sort of pluralist wishes to sort all biological usage under two main umbrella theories, the selected effects theory and the causal role theory. An argument for this sort of pluralism is that it mirrors the two main uses of "function" in biology, the historical sense and the ahistorical sense. If I'm right, this incarnation of the pluralist project can't work either.

True, there are some theories of function I haven't addressed here, which fall a bit outside of the mainstream. Might those come to our rescue? In particular, one might wonder how the *modal theory* of function (Nanay 2010) fares with respect to my analysis. The modal theory holds that a function of a trait *token* depends on that token's behavior on nearby possible worlds, where what's "nearby" depends on our explanatory interests. I agree that this is an ahistorical theory through and through, since what function a trait has, and whether or not it's dysfunctional, depend on what's going on at other possible worlds, rather than the actual past. But it also yields a deeply implausible construal of dysfunction. As Neander and Rosenberg (2012) point out, if the modal theory is right, then many traits that biologists don't think of as dysfunctional, like the trait of lactose-intolerance in most Pacific Islanders, would actually be dysfunctional. So, while the modal theory doesn't violate the *letter* of my second adequacy condition – namely, that it should allow for the possibility of dysfunction – it violates the *spirit* of that condition by carving up functions and dysfunctions in a wildly revisionary way.

Nanay (2012) argues that the fact that function ascriptions are relative to our explanatory interests can somehow lessen the sting of this counterintuitive consequence, but I don't see how that helps. To illustrate the problem, consider Temitope, an evolutionary geneticist who's interested in how human beings might evolve in the near future. Temitope considers a possible world to be "nearby" if, at that possible world, she has a counterpart, and her counterpart's genome differs from hers by only a single point mutation, but the rest of the world is largely the same (yielding at least 3 billion nearby worlds). She reasons that, on some of those possible worlds, some of her traits would do things that enhance her inclusive fitness. For example, we might suppose that there is a possible world at which her body's ability to dissolve arterial plaque is substantially enhanced, one at which she has tetrachromatic vision, and one at which she's resistant to malaria. She realizes, with dismay, that her body's actual ability to dissolve arterial plaque represents a dysfunction. In fact, she realizes that, *relative to her explanatory interests*, she has many more dysfunctions than she ever thought possible. So even if we agree that function ascriptions are tethered to explanatory interests, we still get deeply revisionary consequences. In my reckoning, a theory that hangs together pretty well with ordinary biological usage is better than a deeply revisionary one, all things equal (see Garson 2016, 105-7, for further discussion).

There's a twist to my story, which I alluded to in the introduction. I think there is a prominent sense of "function" in scientific circles that is ahistorical. Consider that climate change is a function of deforestation, poor academic performance is a function of malnutrition, and wildlife habitat is a function of soil. These notions are ahistorical through and through. "Function," in this context, means little more than "effect," and perhaps (as in the last of the three examples) "helpful effect." But this tepid sense of function isn't going to sustain a distinction between function and accident, nor will it give us any sense of dysfunction. This is the sort of "function" that Bock and von Wahlert (1965, 274) were getting at when they equated functions with "all physical and chemical properties arising from [the trait's] form." It's also the sort of "function" that Neander (2017) describes in her recent discussion of "minimal functions." But the proponents of

the allegedly ahistorical theories want functions to do much more than that. They are trying to capture the ordinary biological sense (or *an* ordinary biological sense) of "function," where functions differ from accidents and sometimes things are dysfunctional. Unfortunately, they can't have what they want.

References

Amundson, R., and G. V. Lauder. 1994. Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy* 9: 443-469.

Bigelow, J., and Pargetter, R. 1987. Functions. *Journal of Philosophy* 84: 181-196.

Bock, W. J., and von Wahlert, G. 1965. Adaptation and the form-function complex. *Evolution* 19: 269-299.

Boorse, C. 1976. Wright on functions. *Philosophical Review* 85: 70-86.

Boorse, C. 1977. Health as a theoretical concept. *Philosophy of Science* 44: 542- 573.

Boorse, C. 2002. A rebuttal on functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M. Perlman, 63-112. Oxford: Oxford University Press.

Craver, C. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72: 741–765.

Garson, J. 2016. *A Critical Overview of Biological Functions*. Dordrecht: Springer.

Garson, J. 2018. How to be a function pluralist. *British Journal for the Philosophy of Science* 69: 1101-1122.

Garson, J. 2019. *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press.

Godfrey-Smith, P. 1993. Functions: Consensus without unity. *Pacific Philosophical Quarterly* 74: 196-208.

Griffiths, P. 2009. In what sense does 'nothing make sense except in the light of evolution'? *Acta Biotheoretica* 57: 11-32.

Hardcastle, V.G. 2002. On the normativity of functions. In *Functions: New Essays in the Philosophy of Psychology and Biology*, ed. A. Ariew, R. Cummins, and M Perlman, 144-156. Oxford: Oxford University Press.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Nanay, B. 2010. A modal theory of function. *Journal of Philosophy* 107: 412-431.

Nanay, B. 2012. Function attribution depends on the explanatory context: A reply to Neander and Rosenberg's reply to Nanay. *Journal of Philosophy* 109: 623-627.

Neander, K. 1983. *Abnormal Psychobiology*. Dissertation, La Trobe.

Neander, K. 1991. Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.

Neander, K. 2017. Functional analysis and the species design. *Synthese* 194: 1147-1168.

Neander, K., and Rosenberg, A. 2012. Solving the circularity problem for functions. *Journal of Philosophy* 109: 613-22.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47: 373–388.

Walsh, D.M. 1996. Fitness and function. *British Journal for the Philosophy of Science* 47: 553-574.

Walsh, D. M., and A. Ariew. 1996. A taxonomy of functions. *Canadian Journal of Philosophy* 26: 493-514.

Wright, L. 1973. Functions. *Philosophical Review* 82: 139-168.

Identifying Causes in Psychiatry

Lena Kästner
Ruhr-Universität Bochum, Institut für Philosophie II
mail@lenakaestner.de

Abstract

Explanations in psychiatry often integrate various factors relevant to psychopathology. Identifying genuine causes among them is theoretically and clinically important, but epistemically challenging. Woodward's interventionism appears to provide a promising tool to achieve this. However, it cannot be applied to psychiatry. I thus introduce *difference-making interventionism* (DMI), which detects relevance in general rather than causation. DMI mirrors the empirical reality of psychiatry even more closely than interventionism, but it needs to be supplied with additional heuristics to disambiguate between causes and other difference-makers. To achieve this, I suggest employing heuristics based on multiple experiments, temporal order and scientific domain.

Keywords

causation, psychiatry, interventionism, manipulability, difference-making interventionism, time, scientific domain, multiple realization

word count (incl. references and footnotes): 4674

Paper submitted to: Philosophy of Science Association

1. Causal Explanations in Psychiatry

Causal explanations are abundant in psychiatry. Common statements include, e.g., that traumatic life events cause depressions, that excess dopamine causes mania, that fear causes patients' heart rates to accelerate, and that a spiders crawling on an arachnophobic's arm caused her panic attack. Philosophers traditionally struggle with such causal claims linking processes in the physical domain (dopamine levels, heart rate, a spider's crawling on skin) to those in the realm of the mental (trauma, depression, fear, panic attack). But despite convincing theoretical arguments to the effect that the mental is causally pre-empted by the physical (e.g. Kim 1998), talk of causation is pervasive in the empirical reality of psychiatry.

One may suspect, of course, that causal explanations in psychiatry are merely an instance of "sloppy talk". But while in everyday clinical practice psychiatrists might sometimes use "cause" as shorthand for "difference-maker", psychiatry as a scientific discipline cannot afford to blur the distinction between genuine causes and other difference-makers (background conditions, parts of a whole, realizers, etc.). Building models of psychiatric diseases, disambiguating between alternative explanations, and selecting treatment options all may require understanding what genuine causes of a given condition are. Memory loss after a car crash, for instance, may result from posttraumatic stress disorder (PTSD) or gas poisoning that occurred during the accident. Clearly, the treatment will differ depending on whether poisoning or trauma caused memory loss. Distinguishing causes from other difference-makers in explanations of mental illness thus is not merely a philosophical apprentice piece but highly clinically relevant.

With his manipulability-based *interventionism*, Woodward (2003) proposed an account of causation and causal explanation that aligns nicely with experimental research practice. As such, it seems a promising candidate to overcome (philosophical) worries of causal exclusion and evaluate causal claims in psychiatry (Kendler and Campbell 2009, p. 886). However, interventionism is not up to the task: it is designed to assess causal relations *only* in contexts where no other dependence relations are present. But psychiatrists' causal claims often relate mental (memory loss, depression) and physical (car crash,

neurotransmitters) domains, which—problematically for the interventionist—are usually considered to stand in non-causal dependence relations. No matter whether we characterize cognitive processes as *based* on, *grounded* in, *implemented*, *realized*, *constituted*, or *underlain* by, or *supervening* on neurophysiological process in the brain: as soon as non-causal relations are at issue, interventionism cannot be applied.

This problem is a principled one; it shows up in many special sciences exhibiting a certain inter- or multi-level character. Many modern special sciences combine insights gained at different levels of investigation, from different scientific domains. The same principled questions we are dealing with when relating psychology and neuroscience in psychiatric explanations also arise when linking factors from, say, astronomy, and physics, molecular biology and population genetics, or anthropology and meteorology. But for the time being let us focus on psychiatry.

The current paper offers a systematic account of how to identify causes in psychiatry despite the presence of non-causal dependence relations.[1] I suggest a weakened version of interventionism, *difference-making interventionism* (DMI), that allows applying interventionist methodology despite the presence of non-causal dependence relations. DMI acknowledges that observed manipulability underdetermines the underlying dependence relation. But this loss in specificity is a virtue rather than a vice: Psychiatrists often employ interventions (e.g. in randomized control trials) *before* they know what kind of relation grounds the relevance of difference-makers. Likewise, recent attempts to model psychiatric disorders by means of networks (e.g. Borsboom 2017) incorporate a range of different factors (physiological, genetic, environmental, behavioral), regardless of the possible dependence relations between them. DMI captures and accommodates for this. Besides, it highlights that thinking about explanatory relevance in terms of causal relevance only is too narrow. Still, embracing DMI does neither mean causal relevance loses its special status, nor that we cannot disambiguate between different kinds of dependence relations *in principle*.

[1] This is not a paper on the metaphysics or semantics of causation. I am here primarily concerned with the epistemic question of how to identify causal relations in cases where other dependence relations are present as well, particularly in psychiatry.

Disambiguation between causal and non-causal dependencies might be achieved, I suggest, by drawing on resources other than interventions: We may not only combine evidence from *many experiments* to infer systematic dependencies but can also supplement difference-making graphs with the *dimensions* of *time* and *domain*. A third dimension may accommodate for *multiple realizers*. Thus equipped, DMI preserves the manipulationist core of interventionism while relying on additional heuristics helps identify causes among difference-makers.

Section 2 introduces Woodward's interventionism and the problems with applying it to psychiatry. Section 3 introduces DMI along with some heuristics that may help us identify genuine causes. Section 4 concludes.

2. Woodward's Interventionism and its Application in Psychiatry

According to Woodward's (2003, 2008) *interventionist account of causation*, causal relations can be detected by difference-making: causes make differences to their effects. Causal explanations thus embody a "what-if-things-had-been-different conception of explanation" (Woodward 2003, p. 228). That is, they tell us what will (or would) happen under a range of different circumstances. Inspired by work on causal modeling (Spirtes et al. 1993, Pearl 2000), Woodward translates questions about causal relations into questions about relations between variables (representing properties or events) taking different values. Causal relations between variables can be represented in *directed acyclic graphs*. On the interventionist account, we can infer that X causes Y if and only if we can carry out an intervention on X with respect to Y. A manipulation of X qualifies as an *intervention I on X with respect to Y* if and only if I meets the following conditions: (i) I causes X, (ii) I overrides all other causes of X, (iii) any directed path from I to Y goes through X (i.e. there must not be a causal path, neither direct nor through other variables, from I to Y that does not go through X), and (iv) I is statistically independent of any variable Z that causes Y and that is on a directed path that does not go through X. That is to say, an intervention I

"breaks off" all other influences on X and manipulates X in such a way that changes in Y are *only* mediated through changes in C and *not in any other way*.[2]

Interventionism thus defined mirrors the manipulative character of experimental research practice and reflects certain well-known principles of experimental design (e.g. randomized control trials). It thus allows us to *directly infer* causal relations on the basis of observed manipulability in well-designed empirical studies. Prima facie then, assuming adequate experimental standards apply in psychiatry, we might think that the interventionist view "provides a single, clear empirical framework for the evaluation of all causal claims in psychiatry" (Kendler and Campbell 2009, p. 886; see also Campbell 2007, 2016, Rescorla 2017).

Consider the following toy examples: Tom was mentally healthy before his father died. But as part of his bereavement reaction Tom started grieving. While initial grief after bereavement is not a mental illness, Tom's bereavement experience was so severe and long-lasting that it developed into full-blown depression, although nothing else had changed about his life.[3] Given this picture, it makes sense for the psychiatrist to infer that Tom's grief (more precisely, his bereavement experience characterized by severe grief) caused his depression. Similarly, if Gina's heart rate was at 69bpm while she was relaxing on the sofa before it suddenly accelerated to 132bpm as she was afraid there was a burglar in the hallway (while nothing else changed) it makes sense to infer that Gina's fear of a burglar caused her heart rate to accelerate. And so on. We can picture this with the graphs are shown in figure 1.

[2] Woodward captures this in his definitions (M) and (IV). While (M) expresses that causation is grounded in manipulability relations, (IV) explicates the conditions under which a manipulation is an appropriate intervention, i.e. suitable to uncover causal relations. For the full definitions see Woodward (2003) p. 59 and p. 98, respectively.

[3] The precise relations between bereavement experiences (including grief, apathy, guilt, etc.) and depression are more complex than the current example suggests (see, e.g., Pies 2014, Wagner 2014 for discussions about dropping the bereavement exclusion in DSM V). For current purposes, however, the simplified scenario of Tom developing a depression some time after his bereavement experience, which was primarily characterized by severe grief, will do. For brevity, I shall talk about Tom's grief.

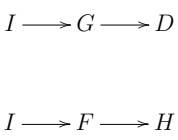


Figure 1: Two independent causal graphs illustrating Gina's and Tom's cases. Setting the value of G (representing Tom's grief) from 0 to 1 changes the value of D (representing his depressive state) from 0 to 1. Setting the value of F (representing Gina's fear) from 0 to 1 changes the value of H (representing her heart rate) from 69 to 132bpm. The interventions (I) are the death of Tom's father and Gina's hearing footsteps in the hallway.

This reasoning squares well with clinical evidence to the effect that grief can cause depressive episodes (e.g. Beck and Alford 2009) and that mental states such as fear affect an individual's heart rate (e.g. Cuthbert et. al. 2003). Likewise, there is evidence for manipulations in the mental domain, like cognitive behavioral therapy (CBT), to be efficacious in treating conditions such as anxiety and depression (e.g. Sofronoff, Attwood, and Hinton 2005, Butler et al. 2006). But should this lead to us conclude that interventionism allows us to establish notoriously difficult mental-to-mental[4] (in Tom's case) as well as mental-to-physical (in Gina's case) causal claims? No. The cases are not as unproblematic for the interventionist as they might seem.

To ensure we can sort actual causes from confounding factors and accidental correlates, interventionist analyses require that all of the considered variables in a causal graph must in principle be *independently manipulable* (see Woodward 2008, p. 209, Woodward 2015). For if this were not the case, we will run risk of violating (iii) and (iv). Thus, if we assume there is some sort of systematic (implementation, realization, supervenience, grounding, …) relation between mental and physical phenomena, applying interventionism is—despite its intuitive plausibility—simply *not licensed* (Eronen 2012, Raatikainen 2010, Shapiro 2010, Shapiro & Sober 2007, Kästner 2017).

Recently, a number of attempts have been made to save interventionism for scenarios with non-causal relations (for discussions in the context of mental causation see, e.g., Woodward 2008, 2015, Baumgartner 2010, 2013, Gebharter

[4] For the sake of the example suppose Tom's depression (at least his depressive mood) is a mental phenomenon.

2015, Hoffmann-Kolss 2014, Kästner 2017; an analogous debate in the context of constitutive mechanistic explanations is reviewed in Kästner and Andersen 2018). The proposed modifications typically advocate either splitting causal graphs or introducing exception-clauses for non-causal dependence relations. But even if these strategies were successful, neither is convincing in the case of psychiatry. First, psychiatrists typically aim at *integrated* explanations relating different (mental, neurophysiological, genetic, …) factors. These factors may be relevant for different reasons: because they exert a causal influence, because they are a part of the implementational (realization, supervenience) base of a certain psychopathology, because they are background conditions, etc. Interventionism, by contrast, is designed to assess *causal relations only*. Second, the exact relations between different variables often remain subject to investigation and cannot be presupposed before scientists start testing for manipulability.

Current network models of mental disorders (e.g. Borsboom 2017, Borsboom, Cramer & Kalis forthcoming) illustrate this. These models are typically based on interventionist reasoning. However, they usually include concrete symptoms along with behavioral, cognitive, genetic, demographic, and environmental factors as variables. While some of these may in fact be related causally (e.g. grief causing depressive mood), for others that seems at least questionable. Is Tom's low serotonin, for instance, implementing or causing depression?[5] And does his socio-economic situation causally contribute or is it merely a background condition? Despite such questions, network models are powerful tools to figure out which factors are relevant to mental disorders. Applying interventionist reasoning to them helps uncover which factors influence one another, as well as what the developmental dynamics are. Moreover, the manipulationist strategy matches well with empirical research reality and provides our best currently available account of scientific (causal) explanation. Thus, it seems well worth trying to save interventionism for psychiatry.

[5] Talking about "low serotonin" as the substrate of depression is probably too simplistic; you might consider it a placeholder for whatever the neurophysiological substrate of depression according to your favorite account.

3. Difference-Making Interventionism for Psychiatry

To save interventionist reasoning for psychiatry, I introduce a weakened form of interventionism: *difference-making interventionism* (DMI). Rather than limiting our analysis to causal relations, DMI identifies a whole bundle of difference-making relations (among them, of course, causal relations). Thus, we can apply DMI in cases where non-causal dependence relations are present. To balance the resulting loss in specificity, we can employ additional heuristics allowing us to identify genuine causes.

3.1 The Bare Bones of DMI

To uncover dependence relations with DMI, we can keep using variables and directed graphs (now speaking of difference-making rather than causal paths) and proceed by the familiar interventionist method: DMI takes X to be a difference-maker for Y with respect to a given variable set V if and only if there is a possible (IV_{dm})-defined intervention on X with respect to Y that will change Y when all other variables Z_i in V are held fixed, except for those on a *difference-making path* from X to Y. (IV_{dm}) defines an intervention as follows: A manipulation I of X qualifies as an *intervention on X with respect to Y* if and only if it meets the following conditions: (i) I is a difference-maker for X, (ii) I overrides all other difference-makers influencing X, (iii) any directed path from I to Y goes through X (i.e. there must not be a difference-making path, neither direct nor through other variables, from I to Y that does not go through X), and (iv) I is statistically independent of any variable Z_i in V making a difference to Y and that is on a directed path that does not go through X.[6]

The advantages of DMI are that we no longer need to worry about restricting our variable set to independently manipulable variables or knowing among which variables non-causal dependencies obtain *before* we proceed to test for manipulability. This is empirically realistic as it reflects that (a) conceiving of explanatory relevance as causal relevance *only* is too narrow and (b) psychiatrists often employ interventionist reasoning (e.g. in randomized

[6] This is structurally analogous to Woodward's definitions (M) and (IV), just modified to no longer restrict our analysis to causal relations (cf. section 2).

control trials) *before* they know what kind of relation grounds the observed difference-making relation. The clinical efficacy of antidepressants, for instance, underdetermines why these drugs work. Do they work because they target the cause of disease or because they interfere with the pharmacological mechanism implementing certain pathologies? Likewise, does CBT help alleviate Tom's depression because it directly targets his mood or because it otherwise induces changes in, say, the low serotonin levels underlying his depression? DMI explicitly acknowledges this underdetermination.

However, the caveat is a significant loss in specificity. Once we adopt DMI, manipulability can no longer be used to directly infer causal relations (otherwise we would face an inflation of causal claims!); DMI underdetermines the underlying dependence relation. But, I propose, this is a virtue rather than a vice: integrating causal and other explanatory factors into a single model is a key feature of network explanations of mental disorders (e.g. Borsboom 2017; Borsboom, Cramer & Kalis forthcoming). Of course, we still need some way or other to identify causes among explanatorily relevant factors. To achieve this, I suggest, we may supplement interventions with other strategies.

3.2. Heuristic Inferences: Asymmetry and Multiple Experiments

It is a platitude about causation that causes are spatiotemporally distinct from their effects (e.g. Lewis 1970); causes precede their effects, and effects depend on their causes but not vice versa. Building on this knowledge, we gain at least two possible criteria to distinguish causation from other forms of dependence: time and asymmetry.[7] Let us first consider asymmetry.

Asymmetric manipulability is a first indication of but not by itself sufficient to infer causation. Take Gina's case: suppose we find that as the footsteps (I_1) induced Gina's fear (F) her heart rate (H) accelerated. But we can also get her heart rate to accelerate if we put her on a treadmill (I_2), which does not induce Gina's fear. This may lead us to infer that F causes H since we can intervene into F with respect to H but not vice versa. However, H may also be a supervenience

[7] Both strategies only work so long as we do not commit to simultaneous causation. Feedback loops can, however, be accommodated for in terms of repeated causal interactions between the same factors at different points in time (A causes B at t1 and B causes A at t2, …) once we take into account temporal order and draw out feedback loops over time (see section 3.3).

base or realizer of F such that changes in F are necessarily accompanied by changes in H while *only some* changes in H will be accompanied by changes in F (and for the treadmill it was not).

We can thus derive the following heuristic for identifying causes *across multiple experiments* or repetitions: provided that repeated interventions on F with respect to H do affect H, we should consider F a genuine cause of H when a critical number of interventions (say, 1.000) into H with respect to F fails. If, by contrast, some of these interventions into H with respect to F actually affect F, it seems more plausible that F supervenes on / is realized by H.[8]

3.3. Adding Dimensions: Time, Domain and Multiple Realizers

Let us turn to time. Causation is typically considered *diachronic* (see fn 7) while non-causal dependence relations like realization, implementation, constitution, supervenience, part-whole, etc. are usually considered *synchronic* in nature. Thus, the *temporal profile* of variables changing their values in response to a given intervention may give us a clue as to whether or not variables are causally related. Notice, however, that inferences based on temporal order may be compromised by practical and methodological constraints. How quickly can my measurement technique detect changes? What are adequate timescales to consider (generations, hours, nanoseconds, …)? And how should I individuate variables to begin with?

While variable individuation is a notorious problem for the interventionist (and constraints will likely depend on the purpose of our analysis as well as the nature of the scenario) time often is already implicit in how we draw causal graphs: from left to right, starting with variables representing early occurrences of properties or events. Yet, what matters for interventionists merely is which variables are linked through edges, not how they are positioned. By adding an "arrow of time" into the picture, respecting the order in which variables take their values, and including variables changing their values over time as multiple variables (see also Gebharter and Kaiser 2014), we can make the temporal dimension explicit (see figure 2). Despite the possible increase in

[8] Analogous suggestions are made by Baumgartner and Gebharter (2016) and Baumgartner and Casini (forthcoming) to identify mechanistic constitution.

complexity, this adds quite some representational power to difference-making graphs while helping us use temporal order to distinguish different relevance relations.

In addition to temporal aspects, *scientific domain* can also give us useful clues as to what relations might mediate observed manipulability. The idea is perhaps best described by reference to the familiar levels-metaphor. Variables representing properties or events from the cognitive (mental, psychological) domain are usually regarded as on a "higher" level than, say, variables representing properties or events from the "lower"-level neurophysiological or genetic domains. My use of "levels" here is not tied to any specific account of levels; neither do I want to impose any specific hierarchy of domains. What matters for current purposes is merely whether two variables are *located in the same scientific domain*.[9] If they are not, we can draw on *systematic knowledge* (or assumptions we may have) about how variables from the domains in question relate and project that into our graphs.

For instance, cognitive processes are usually considered as neurophysiologically implemented by neural processing in the brain. Accordingly, variables representing mental processes, e.g. memory or depression, should be assumed to relate to variables representing, e.g., low serotonin or hippocampal long-term potentiation (LTP) by implementation rather than causation. Similarly, insights about *specific* containment relations (this AMPA receptor is located in the postsynaptic membrane of hippocampal CA1) can be used to distinguish difference-making mediated by causal dependencies from difference-making due to part-whole relations. Graphically, we can represent such insights by placing variables in our graphs along a vertical dimension and marking known relations with specific kinds of arrows (see figure 2).

In principle, heuristic inferences based on time and domain should be considered independent, neither is primary. Insights on time and domain are

[9] For current purposes, I am using the term "scientific domain" rather non-technically to refer to scientific fields or areas of research like neuroscience, psychology, astronomy, or genetics. However, scientific domains may cross-cut layers of a traditional layer cake picture of science and how to best individuate scientific domains may depend on the research project at hand (see Kästner 2018).

usually acquired in different ways: Information on time is typically gathered through intervention-based studies. Knowledge—or assumptions—about systematic relations between domains or specific containment relations, by contrast, is usually acquired through meta-scientific reasoning or theorizing as well as non-intervention studies. Non-intervention studies do not manipulate some factor X with respect to another factor Y but study features like structure and organization of a system by other means (see Kästner 2015). Examples include staining, tracing, cutting-open, centrifuging, and x-raying. Moreover, insights on time and domain tend to play different roles in our search for causal relations: The temporal order revealed by interventions tends to suggest candidate causal relations. Information about relations between concrete variables or between variables from different domains typically restricts which manipulability relations might be considered candidates for causal relations. At times, the two strategies may constrain one another or deliver conflicting results. When conflicting results occur, we must decide which one to prioritize based on the reliability of the information we fed into our heuristics to begin with.

Finally, we might take into account multiple realizers to acknowledge that, e.g., mental disorders may be neurophysiologically implemented in different ways. Depending on how a certain psychopathology is realized in a patient it may develop and be influenced in different ways. This is important to understand, say, why some patients respond well to certain treatments and others do not. Graphically, we can capture this in different planes along a third dimension of difference-making graphs where the structure of the graph may differ between planes.

Visualizing our insights from these different heuristics in difference-making graphs supplied with dimensions and multiple kinds of arrows helps us construct integrative network explanations in psychiatry without loosing sight of actual causal relations. Figure 2 illustrates what this may look like for Tom's case.

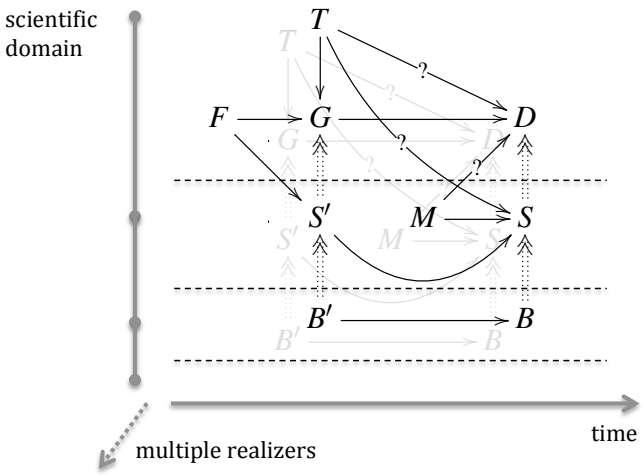


Figure 2: A difference-making graph of Tom's case. CBT (T) influences Tom's grief (G) and possibly his depressive state (D) as well as his serotonin levels (S). Gene expression as measured using biomarkers (B', B) may be considered the implementation base or realizer (dotted double arrows) of serotonin levels (S', S). Medication (M) targets serotonin levels but may also have a cognitive effect (possibly a placebo-effect) on D. Different kinds of relations are marked with different kinds of arrows and variables are arranged along three dimensions (time, domain, realizers). The grey graph symbolizes the inclusion of multiple realizers in different planes; causal relations may differ between these planes.

4. Conclusions

Causal explanations in psychiatry are not merely a result of "sloppy talk"; distinguishing between causes and other difference-makers is actually highly relevant for psychiatry as a scientific discipline as well as for clinical practice. However, multiple different factors may contribute to a given psychopathology in various ways. Thus restricting our analysis to causal relations *only* is too limited. DMI acknowledges this by modifying the Woodwardian interventionism such that it can be applied in contexts where non-causal dependencies are present. Still, identifying causal relations among other dependencies remains an epistemically demanding endeavor. Employing additional heuristics based on multiple experiments, considerations of time and scientific domain may help us identify genuine causes among difference-makers.

Acknowledgements

I am grateful to Astrid Schomäcker, Sanneke de Haan, Henrik Walter, Juan Loaiza, Dimitri Coelho Mollo, Michael Pauen and two anonymous reviewers for comments on an earlier version of this manuscript. Many thanks also to Jon Williamson, Albert Newen, and the members of the Philosophy of Psychiatry Reading Group at Berlin School of Mind and Brain for discussions on the matter.

References

Baumgartner, M. (2010). Interventionism and epiphenomenalism. Canadian Journal of Philosophy, 40 (3), 359–383.

Baumgartner, M. (2013). Rendering interventionism and non-reductive physicalism compatible. Dialectica, 67, 1–27.

Baumgartner, M. & Casini, L. (forthcoming). An Abductive theory of constitution. Philosophy of Science.

Baumgartner M. & Gebharter, A. (2016). Constitutive relevance, mutual manipulability, and fat-handedness, British Journal for the Philosophy of Science, 67, 731–756.

Beck, A.T. & Alford, B.A. (2009). *Depression: Causes and Treatment*. Philadelphia, PA: University of Pennsylvania Press.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, 16, 5–13.

Borsboom, D., Cramer, A. & Kalis, A. (forthcoming). Brain disorders? Not really... Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*.

Butler, A.C., Chapman, J.E., Forman, E.M. & Beck, A.T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review*, 26, 17–31.

Campbell, J. (2007). An interventionist approach to causation in psychology. In: A. Gopnik & L. Schulz (eds.), Causal Learning: Psychology, Philosophy, and Computation. New York: Oxford University Press, pp. 58–66.

Campbell, J. (2016). Validity and the Causal Structure of a disorder. In: Kendler, K. and Parnas, J. (eds.) Philosophical Issues in Psychiatry IV: Psychiatric Nosology. Oxford: Oxford University Press.

Cuthbert, B.N., Lang, P.J., Strauss, C., Drobes, D., Patrick, C.J. & Bardley, M. (2003). The psychophysiology of anxiety disorder: Fear memory imagery. *Psychophysiology*, 40, 407-422.

Eronen, M. I. (2010). Reduction in Philosophy of Mind: a Pluralistic Account. Ph.D. thesis, University of Osnabrück.

Eronen, M. I. (2012). Pluralistic physicalism and the causal exclusion argument. European Journal for Philosophy of Science, 2, 219–232.

Gebharter, A. (2015). Causal exclusion and causal Bayes nets. Philosophy and Phenomenological Research. DOI:10.1111/phpr.12247.

Gebharter, A. & Kaiser, M. I. (2014). Causal graphs and biological mechanisms. In: M. I. Kaiser, O. Scholz, D. Plenge, & A. Hüttemann (eds.), *Explanation in the special sciences: The case of biology and history*, Dordrecht: Springer, pp. 55–86.

Hoffmann-Kolss, V. (2014). Interventionism and Higher-Level Causation. International Studies in the Philosophy of Science, 28, 49-64.

Kästner, L. & Andersen, L. (2018) Intervening into Mechanisms: Prospects and Challenges. *Philosophy Compass*. Manuscript under revision.

Kästner, L. (2015). Learning About Constitutive Relations. In: U. Mäki, S. Ruphy, G Schurz & I. Votsis (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki,* pp. 155-167. Springer.

Kästner, L. (2017). *Philosophy of Cognitive Neuroscience: Causal Explanations, Mechanisms & Experimental Manipulations*. Berlin: Ontos/DeGruyter.

Kendler, K.S. & Campbell, J. (2009). Interventionist casual models in psychiatry: repositioning the mind-body problem. Psychological Medicine, 39, 881887.

Kim, J. (1998). Mind in a Physical World. Cambridge: MIT Press.

Lewis, D. (1970). Causation. Journal of Philosophy, 70, 556—567.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press.

Pies, R.W. (2014). The bereavement exclusion and DSM-5: An update and commentary. *Innovations in Clinical Neuroscience*, 11, 19–22.

Raatikainen, P. (2010). Causation, exclusion, and the special sciences. Erkenntnis, 73, 349–363.

Rescorla, M. (2017). An interventionist approach to psychological explanation. *Synthese*. DOI: 10.1007/s11229-017-1553-2

Shapiro, L.A. (2010). Lessons from causal exclusion. Philosophy and Phenomenological Research, 81, 594–604.

Shapiro, L. & Sober, E. (2007). Epiphenomenalism - The do's and the don'ts. In: P. Machamer & G. Wolters (eds.), Thinking about Causes. Pittsburgh: University of Pittsburgh Press. pp. 235–264.

Sofronoff, K. Attwood, T. & Hinton, S. (2005). A randomised controlled trial of a CBT intervention for anxiety in children with Asperger syndrome. *Journal of Child Psychology and Psychiatry*, 46, 1152-1160.

Spirtes, P., Glymour, C. & Scheines, R. (1993). Cauation, Prediction and Search. New York: Springer.

Wagner, B. (2014). *Komplizierte Trauer*. Berlin und Heidelberg: Springer.

Woodward, J. (2003). Making Things Happen. Oxford: Oxford University Press.

Woodward, J. (2008). Mental causation and neural mechanisms. In: J. Hohwy & J. Kallestrup (eds.), Being Reduced: New Essays on Reduction, Explanation, and Causation. Oxford: Oxford University Press, pp. 218–262.

Woodward, J. (2015). Interventionism and causal exclusion. Philosophy and Phenomenological Research, 91, 303-347.

The Reference Class Problem for Credit Valuation in Science

Carole J. Lee (c3@uw.edu)

Abstract: Scholars belong to multiple communities of credit simultaneously. When these communities disagree about a scholarly achievement's credit assignment, this raises a puzzle for decision and game theoretic models of credit-seeking in science. The reference class problem for credit valuation in science is the problem of determining to which of an agent's communities – which reference class – credit determinations should be indexed for any given act under any given state of nature. I will identify strategies and desiderata for resolving ambiguity in credit valuation due to this problem and explain how pursuing its solution could, ironically, lead to its dissolution.

1. Introduction

Within the scientific community, there is a common understanding that its reward system drives problematic behavior linked to publication patterns, pipeline retention, hypercompetitive scientific cultures, and reproducibility. Conversely, there is also a shared sentiment that, in order to change these cultures and behaviors in ways that would improve science, the scientific community must coordinate across institutions to change how credit is assigned at the level of the individual scientist (Alberts et al. 2014, Nosek et al. 2015, Aalbersberg et al. 2017, National Academies of Sciences 2018, National Science Foundation 2015, Blank et al. 2017). The hope is

that increasing individual researchers' incentives towards increased transparency and openness will improve the integrity, reproducibility, and accuracy of the published record.[1]

Analogously, philosophers working in the "credit economy" tradition adopt the working assumption that there is some amount of credit that agents can accrue for different acts under different states of nature. This assumption allows them to use decision and game theoretic tools to model how credit-seeking among individual scientists can give rise to behavior and norms that support or thwart the achievement of community-wide goals. When, in the aggregate, individual credit-seeking cuts against collective ends, their approach can explore how changes to individuals' incentive structures can nudge and redirect individual behavior (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Bright 2017, Heesen 2017, Kitcher 1990, Strevens 2003, Zollman 2018). Different philosophers make different assumptions about the norms by which credit gets allotted – for example, whether credit is best thought of as all-or-nothing (Strevens 2003, Bright 2017, Heesen 2017) or as something that may come in degrees (Bruner and O'Connor 2017, Rubin and O'Connor 2018, Zollman 2018). However, the general approach assumes that there is some precise way to assign credit to different acts under different states of nature – an assumption that allows these philosophers to model credit-seeking behavior and the emergence of scientific norms in formally tractable ways.

But, how much credit gets assigned to any given act under any given state of nature? Just as each of us simultaneously belongs to multiple social categories, each of which is tied to implied social hierarchies (Macrae, Bodenhausen, and Milne 1995, Crenshaw 1989), each

[1] Institutions can also experience incentives that promote or thwart scientific ends (Lee and Moher 2017).

scholar simultaneously belongs to multiple communities of value with implied social hierarchies for assigning credit. To which of an agent's communities – which reference class – should credit determinations be indexed and why?

In this paper, I will use examples from the current context of science's complex and dynamic culture to motivate and illuminate what I will call the *reference class problem for credit valuation in science*. I will identify a few strategies and desiderata for solving ambiguity in credit assignments due to the reference class problem. And, I will say a bit about how developing the resources needed to solve it could ultimately sow the seeds for its own dissolution.

2. The Reference Class Problem for Credit Valuation in Science

The contours of this puzzle about the "coin of recognition" (Merton 1968, 56) become visible when one moves beyond thinking about credit in generic, abstractions of scientific communities towards the heterogeneous communities we find today. I start from this more concrete perspective because prestige requires recognition *by individuals and forums* that are themselves valued by credit-seeking scholars (Zuckerman and Merton 1971, Lee 2013): credit worthiness in science is a function of the individuals and systems designed to assess, allocate, dispute, and enforce it. Although some aspects of Zuckerman and Merton's narrative about the origins of the normative structure of science have been contested by historians (Csiszar 2015, Biagioli 2002), we see the social dynamics Zuckerman and Merton proposed clearly at play in contemporary science. For example, Nature Publishing Group recently found that – for the 18,354 authors in science, engineering, and medicine surveyed – the reputation of a journal is the primary factor driving choices about where to submit their work, where reputation is primarily

determined by the journal's impact factor and its standing "as the place to publish the best research" (Nature Publishing Group 2015).[2] Factors associated with a journal's ability to archive and disseminate research – things like a journal's time from acceptance to publication, indexing services, or Open Access options – are much less important.[3]

Within academia, each of us simultaneously belongs to multiple communities of value. The reference class problem arises when these different communities of value disagree about the amount of credit an agent accrues for different acts under different states of nature. Although I take this problem to be general, for the sake of clarity and simplicity in presentation, I will focus my examples on communities that can be described as having a nesting structure: for example, individual scholars belong to specific sub-disciplines, which are nested within disciplines, which are nested within a more general population of scholars. A sub-population that is nested within a population can have a credit sub-culture whose valuations differ from that of the population, whose valuations can differ from that of the super-population. In these cases, changing how

[2] Note that using journal impact factor to measure an individual article's importance is both old-fashioned and problematic: citation distributions within journals are so skewed that it is statistically improper to infer the impact of an individual article on the basis of the impact factor of the journal in which it is published (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017, Larivière et al. 2016, Wilsdon et al. 2015).

[3] Some decision theorists, especially those working outside of philosophy, may reject or remain agnostic about attributing mental states such as beliefs to agents (Okasha 2016). However, because I understand credit and credit-seeking as sociological phenomena involving status beliefs, I am committed to attributing beliefs to agents.

narrowly or broadly one draws the boundaries of an agent's community of valuation can change the amount of credit assigned to a scholarly accomplishment, just as changing how one gerrymanders the boundaries of a voting district can change its election outcomes. This gives rise to the *reference class problem for credit valuation in science*: to which of the agent's communities – which reference class – should credit valuations be indexed when determining the amount of credit the agent accrues for different acts under different states of nature?

There are many examples across academia where nesting community structures can give rise to paradoxes and pathologies in credit assignments. For example, a scholar's individual sense of what counts as quality work – their individual credit assignments – may deviate from what is endorsed in their sub-discipline's or discipline's status hierarchy (Correll et al. 2017, Centola, Willer, and Macy 2005, Willer, Kuwabara, and Macy 2009). A puzzle that has cachet in a sub-discipline may be of peripheral importance within that discipline: for example, a more accurate technique for measuring how temperature cools with elevation considered critical in mountain meteorology and mountain ecology (Mindner, Mote, and Lundquist 2010) may have less visibility, despite its relevance, to the larger discipline of hydrology (Livneh et al. 2013).[4] A question, technique, or approach that is thought to have high impact *across* fields may have less prominence *within* each of those fields. For example, consider a hypothetical scenario involving an interdisciplinary project whose authors and content represent a set of non-overlapping

[4] Indeed, savvy scholars can rebel against their field's disciplinary and sub-disciplinary boundaries to form an "unruly alliance" as a new field, as in the example of solid state physics, which was formed principally to serve the interests of applied physicists by linking their work to related abstract physical research within a new sub-discipline (Martin 2018, 199).

disciplines. Let's imagine that scholars in each of these disciplines prefer purely disciplinary projects over the interdisciplinary project; however, when these scholars' preferences are aggregated, their collective preference is for the interdisciplinary project over any single purely disciplinary project (because they prefer interdisciplinary projects over purely disciplinary projects that originate from outside their own fields). Imagine now that this project gets published in a journal, valued by those disciplines, that seeks papers of interest *across and beyond disciplines* (not just within disciplines) – this is one way to interpret, for example, *Science's* mission to publish papers that "merit recognition by the wider scientific community and general public. . . beyond that provided by specialty journals" (Science).[5] Which reference class would be most relevant in evaluating the value of the interdisciplinary project (and why)?

There are other ways of dividing scholarly communities into nesting structures that create tensions in credit assignments. The pressures a scholar may feel from the incentive structure impacting her department/school may be slightly different from the incentive structure impacting her university. A coarse but concrete way to see this is to think about the prestige structure reified and reinforced by ranking systems (Espeland and Sauder 2012, 2016, Sauder and Espeland 2006), which transform "the ways professional opportunities are distributed" within organizations (Espeland and Sauder 2016, 7). Imagine that an untenured business school professor with a potentially high impact manuscript needs to burnish her prestige in the eyes of

[5] Note that, at its inception in 1869, *Nature* also aimed to share scientific advances "of general interest" with working scientists and the general public (Nature 1869); and, as early as 1893, scholars saw *Nature* as a place where they could reach audiences "across increasingly sharp disciplinary boundaries" (Baldwin 2015, 72).

both her dean and her provost, since both will evaluate her tenure case. If her provost is working to gain stature on the Academic Rankings of World Universities [ARWU], the professor should submit her manuscript to *Science* or *Nature*, since the ARWU ranks universities by their publications in these journals (Academic Ranking of World Universities 2018). However, if her dean is trying to gain stature on the *Financial Times* International ranking of MBA programs, she should submit to one of the fifty business, economics, or psychology journals by which the FT ranking system evaluates business school prestige – notably, the journal list does not include *Science* or *Nature* (Ormans 2016). What should the business school professor do?

Finally, credit assignments can vary depending on how long a time window a scholar keeps in view. A coarse but concrete way to think about this is by looking at how metrics for evaluating scholarship change over time. Journal impact factors are becoming less useful measures for evaluating an individual's scholarly contribution: since the advent of the digital age, the most elite journals (including *Science* and *Nature*) are publishing a decreasing percentage of the top cited papers (Larivière, Lozano, and Gingras 2013); the relationship between journal impact factor and paper citations has declined over time (Lozano, Larivière, and Gingras 2012); and, the citation distributions between journals "overlap extensively" (Larivière et al. 2016). The current wisdom is that if quantitative indicators are to be used to evaluate research, it is more useful to use article-level metrics such as citations as well as alternative metrics such as downloads and views (San Francisco Declaration on Research Assessment 2013, Hicks and Wouters 2015, Wilsdon et al. 2017). On the horizon, there are now calls for creating new metrics that can encourage researchers and journals to be transparent and open in their reporting practices (National Academies of Sciences 2018, Wilsdon et al. 2017, Aalbersberg et al. 2017), where the rise of such metrics – as well as the growing meta-research literature that

ranks journals by the replicability (Schimmack 2015) or sample size and statistical power of their published results (Fraley and Vazire 2014) – makes it possible for a journal's impact factor and epistemic credibility to come apart (Fang and Casadevall 2011). Analogously, these new metrics, if assigned to individual researchers, may also reveal ways in which traditional markers of prestige (e.g., journal impact factor, citations, institutional rank) and epistemic credibility can also come apart. Other dynamic considerations can also give rise to the reference class problem: for example, the community to which a junior scholar aims their accomplishments (e.g., related to hiring within a disciplinary department or professional school) may be different from the audience they wish to command as a senior scholar.

Decision theorists and game theorists capture the risky nature of individual choices by allowing for uncertainty about which states of the world will come to be; and, when the probabilities attached to different outcomes are understood subjectively, these models permit a kind of subjectivity in estimates of expected credit for different acts. However, I hope the examples throughout this section animate genuine *ambiguity in credit* due to the reference class problem for credit valuation in science.

3. Strategies and Desiderata for Solving the Reference Class Problem

How might decision theorists and game theorists try to solve the reference class problem? One possible approach argues for the "correctness" of using one community rather than another.[6]

[6] Note that indexing credit valuation to a particular community need not prevent scholars from outside that community from understanding the relative value of that contribution: for example, if one were to adopt the old-fashioned and problematic assumption that an article's impact can be

For example, it might be tempting to argue that all prestige is discipline-based since many scholarly prizes are distributed for excellence in particular disciplines (e.g., Nobel prize, Fields prize, academic society prizes); and, even when research is funded or published in interdisciplinary contexts, it may be primarily evaluated on the basis of its disciplinary excellence (Lamont 2009, but see Lee et al. 2013). Because this strategy for addressing the reference class problem relies heavily on identifying the "right" community, defending the centrality of the chosen community as opposed to others is critical. For example, some may challenge the idea that disciplines should be the sole arbiter of credit. After all, the awarding of some scientific prizes reach across disciplinary conceptions of excellence (e.g., consider winners of the MacArthur Genius Prize and the psychologists who have won the Nobel Prize in Economics).

Another possible approach creates an algorithm that calculates the credit value of a scholarly contribution by summing the credit valuation of multiple communities. This approach would need to identify exactly how much to weight each community's valuation – with a rationale for why – since different weightings could lead to different overall credit valuations.[7]

---

measured by the impact factor of the journal in which it is published, and one recognizes that citation rates vary across disciplines, one could use field-normalized percentiles to understand a paper's impact in a metric that is legible across fields (Hicks and Wouters 2015).

[7] On the face of it, this may seem like a form of commensuration because it involves summing values to calculate an overall score (Espeland and Stevens 1998). Note that the process of commensuration requires combining values across *qualitatively* different domains of value. As such, this would only count as commensuration if we moved to a pluralistic account involving

Note that some scholars take this style of approach when trying to measure the relative prestige of journals: in particular, the Eigenfactor score rates journals according to the number of its incoming citations, where the "relative importance" of each incoming citation is contextualized by the frequency with which the citing journal is itself cited (West, Bergstrom, and Bergstrom 2010).

Those who may wish to model the implications of different approaches for solving the reference class problem may try to do so by setting up hypothetical communities that assign community boundaries and credit assignments in *de facto* ways to see what kinds of behaviors and norms emerge. This work could reveal interesting insights into how different ways of gerrymandering intellectual populations – by shifting sub-discplinary and disciplinary lines, journal scope, and grant agency program areas/panels – could change the kinds of projects and areas that "win."

However, to solve the underlying *conceptual* problem, one must provide theories of community and credit that address two fundamental but vexing questions. How should one define and gerrymander the boundaries of the relevant communities invoked in the proposed solution? And, how does one determine the amount of credit those communities would assign to different acts under different states of nature? These questions may not be independently answerable. The boundaries of a community may need to be defined in terms of patterns of shared lore among its members about how credit is accrued – shared beliefs that coordinate credit-seeking and enforcement behavior in cases where status beliefs are internalized as norms

summing heterogeneous kinds of credit. For a more straightforward example of commensuration in scientific evaluation, see Lee (2015).

(Merton 1973) and in cases where they are not (Willer, Kuwabara, and Macy 2009, Ridgeway and Correll 2006). Conversely, in recognition that some community members can have more influence than others on the content of reigning status beliefs, a community's credit assignments may need to be defined with some reference to the causal patterns of interaction among specific individuals and clusters of individuals – including status judges who wield "social control through their evaluation of role-performance and their allocation of rewards for that performance" (Zuckerman and Merton 1971, 66). However, answers to these questions should not *exclusively* inform each other. In particular, we must be careful not allow the size of a scholarly population and/or the power of its status judges to fully determine the intellectual value of the questions pursued by any particular partition of the scholarly universe.

4. Conclusion

Scientific credit – the "coin of recognition" (Merton 1968, 56) – is assessed, allocated, disputed, and enforced by many different communities and institutions within science that support and sustain a multiplicity of status hierarchies. This gives rise to what I have called the reference class problem for credit valuation in science. Solving this problem requires developing rich theories of community and credit that are based on fine-grained information about the structure and status systems of complex scholarly networks.

The irony of this assessment is that such investigation towards solving the reference class problem could ultimately sow the seeds for its own dissolution. In particular, such study can render friable a critical assumption for both the reference class problem and for decision theory models: namely, that communities, once defined, assign determinate amounts of monistic credit for different acts under different states of nature – that credit "can vary quantitatively but not

qualitatively" (Anderson 1993, xii).[8] Contrary to this, recent policy papers call for moving away from narrowly conceived measurements of research excellence towards broader ones that are sensitive to the diversity of research missions among individual researchers, programs, and academic institutions (Hicks and Wouters 2015, Wilsdon et al. 2015). Such work can include community-engaged scholarship that creates, disseminates, and implements knowledge in coordination with the public to identify social interventions, change social practice, and influence policy (Hicks and Wouters 2015, San Francisco Declaration on Research Assessment 2013, Boyer 1990, Escrigas et al. 2014). From the perspective of these efforts, plurality in our notions of scholarly excellence and credit – and differences in valuation and prioritization practices between individuals and communities – may be best conceived, not as a logical problem to solve, but as a starting point for theorizing.

Acknowledgments: Many thanks to Christopher Adolph, Melinda Baldwin, Alex Csiszar, Aileen Fyfe, Sheridan Grant, Crystal Hall, Remco Heesen, Liam Kofi-Bright, Jessica Lundquist, Joseph Martin, Conor Mayo-Wilson, Ties Nijssen, Cailin O'Connor, and Kevin Zollman for illuminating conversations. This work is based upon work supported by the National Science Foundation under Grant No. 1759825. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect

[8] Note too that, for formal reasons, the assumption that individual credit assessments could be aggregated into a collective one is questionable given the challenges of combining individual preferences into collective ones (Arrow 1950).

the views of the National Science Foundation. This research used statistical consulting resources provided by the Center for Statistics and the Social Sciences, University of Washington.

References

Aalbersberg, IJsbrand Jan, Tom Appleyard, Sarah Brookhart, Todd Carpenter, Michael Clarke, Stephen Curry, Josh Dahl, Alex DeHaven, Eric Eich, Maryrose Franko, Len Freedman, Chris Graf, Sean Grant, Brooks Hanson, Heather Joseph, Véronique Kiermer, Bianca Kramer, Alan Kraut, Roshan Kumar Karn, Carole Lee, Aki MacFarlane, Maryann Martone, Evan Mayo-Wilson, Marcia McNutt, Meredith McPhail, David Mellor, David Moher, Alison Mudditt Mudditt, Brian Nosek, Belinda Orland, Tim Parker, Mark Parsons, Mark Patterson, Solange Santos, Carolyn Shore, Dan Simons, Bobbie Spellman, Jeff Spies, Matt Spitzer, Victoria Stodden, Sowmya Swaminathan, Deborah Sweet, Anne Tsui, and Simine Vazire. 2017. "Making science transparent by default; Introducing the TOP Statement." *OSF Preprints*. doi: https://doi.org/10.31219/osf.io/sm78t.

Academic Ranking of World Universities. 2018. "ShanghaiRanking's Academic Ranking of World Universities 2018 Press Release." accessed September 1. http://www.shanghairanking.com/Academic-Ranking-of-World-Universities-2018-Press-Release.html.

Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. 2014. "Rescuing US biomedical research from its systematic flaws." *Proceedings of the National Academy of Sciences* 111 (16):5773-7.

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Arrow, Kenneth J. 1950. "A difficulty in the concept of social welfare." *Journal of Political Economy* 58 (4):328-46.

Baldwin, Melinda. 2015. *Making Nature: The History of a Scientific Journal*. Chicago: University of Chicago Press.

Biagioli, Mario. 2002. "From Book Censorship to Academic Peer Review." *Emergences: Journal for the Study of Media & Composite Cultures* 12 (1):11-45.

Blank, Rebecca, Ronald J. Daniels, Gary Gilliland, Amy Gutmann, Samuel Hawgood, Freeman A. Hrabowski, Martha E. Pollack, Vincent Price, L. Rafael Reif, and Mark S. Schlissel. 2017. "A new data effort to inform career choices in biomedicine." *Science* 358 (6369):1388-9.

Boyer, Ernest L. 1990. *Scholarship Reconsidered*. San Francisco, CA: The Carnegie Foundation for the Advancement of Teaching.

Bright, Liam Kofi. 2017. "On Fraud." *Philosophical Studies* 174:291-310.

Bruner, Justin, and Cailin O'Connor. 2017. "Power, Bargaining, and Collaboration." In *Scientific Collaboration and Collective Knowledge*, edited by Thomas Boyer-Kassem, Conor Mayo-Wilson and Michael Weisberg, 135-157. Oxford, UK: Oxford University Press.

Centola, Damon, Robb Willer, and Michael Macy. 2005. "The emporer's dilemma: A computational model of self-enforcing norms." *American Journal of Sociology* 110 (4):1009-40.

Correll, Shelley J., Cecilia L. Ridgeway, Ezra W. Zuckerman, Sharon Jank, Sara Jordan-Bloch, and Sandra Nakagawa. 2017. "It's the conventional thought that counts: How third-order inference produces status advantage." *American Sociological Review* 82 (2):297-327.

Crenshaw, Kimberle. 1989. "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics." *University of Chicago Legal Forum* 139:139-168.

Csiszar, Alex. 2015. "Objectivities in Print." In *Objectivity in Science: New Perspectives from Science and Technology Studies*, edited by Flavia Padovani, Alan Richardson and Jonathan Y. Tsou, 145-69. Cham, Switzerland: Springer International Publishing.

Escrigas, Cristina, Jesús Granados Sánchez, Budd Hall, and Rajesh Tandon. 2014. "Editor's introduction. Knowledge, engagement and higher education: Contributing to social change." In *Report: Higher Education in the World*, edited by Cristina Escrigas, Jesús Granados Sánchez, Budd Hall and Rajesh Tandon. Palgrave Macmillan.

Espeland, Wendy Nelson, and Michael Sauder. 2012. "The Dynamism of Indicators." In *Governance by Indicators: Global Power through Quantification and Rankings*, edited by Kevin Davis, Angelina Fisher, Benedict Kingsbury and Sally Engle Merry, 86-109. Oxford: Oxford University Press.

Espeland, Wendy Nelson, and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York, NY: Russell Sage Foundation.

Espeland, Wendy Nelson, and Mitchell L. Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24:313-43.

Fang, Ferric C., and Arturo Casadevall. 2011. "Retracted Science and the Retraction Index." *Infection and Immunity* 79 (10):3855-9.

Fraley, R. Chris, and Simine Vazire. 2014. "The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power." *PLOS ONE* 9 (10):e109019. doi: 10.1371/journal.pone.0109019.

Heesen, Remco. 2017. "Communism and the Incentive to Share in Science." *Philosophy of Science* 84:698-716.

Hicks, Diana, and Paul Wouters. 2015. "The Leiden manifesto for research metrics." *Nature* 520:429-31.

Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* LXXXVII (1):5-22.

Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Larivière, Vincent, Véronique Kiermar, Catriona J. MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. 2016. "A simple proposal for the publication of journal citation distributions." *BioRxiv*:062109.

Larivière, Vincent, George A. Lozano, and Yves Gingras. 2013. "Are elite journals declining?" *Journal of the Association for Information Science and Technology* 65 (4):649-55.

Lee, Carole J. 2013. "The limited effectiveness of prestige as an intervention on the health of medical journal publications." *Episteme* 10 (4):387-402.

Lee, Carole J. 2015. "Commensuration bias in peer review." *Philosophy of Science* 82:1272-83.

Lee, Carole J., and David Moher. 2017. "Promote Scientific Integrity via Journal Peer Review." *Science* 357 (6348):256-7.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in peer review." *Journal of the American Society for Information Science and Technology* 64 (1):2-17.

Livneh, Ben, Eric A. Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M. Andreadis, Edwin P. Maurer, and Dennis P. Lettenmaier. 2013. "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions." *Journal of Climate* 26 (23):9384-9392.

Lozano, George A., Vincent Larivière, and Yves Gingras. 2012. "The weakening relationship between the Impact Factor and papers' citations in the digital age." *Journal of the American Society for Information Science and Technology* 63 (11):2140-45.

Macrae, C. Neil, Galen V. Bodenhausen, and Alan B. Milne. 1995. "The Dissection of Selection in Person Perception: Inhibitory Processes in Social Stereotyping." *Journal of Personality and Social Psychology* 69 (3):397-407.

Martin, Joseph D. 2018. *Solid State Insurrection: How the Science of Substance Made American Physics Matter*. Pittsburgh, PA: University of Pittsburgh Press.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* 1968:56-63.

Merton, Robert K. 1973. "The normative structure of science." In *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Norman W. Storer, 267-78. Chicago, IL: University of Chicago Press.

Mindner, Justin R., Philip W. Mote, and Jessica D. Lundquist. 2010. "Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains." *Journal of Geophysical Research: Atmospheres* 115. doi: https://doi.org/10.1029/2009JD013493.

National Academies of Sciences, Engineering, and Medicine,. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, D.C.: The National Academies Press.

National Science Foundation. 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. In *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*.

Nature. 1869. "Nature: A Weekly Illustrated Journal of Science." *Nature* 1 (2).

Nature Publishing Group. 2015. "Author Insights 2015 Survey."

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Mahlhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. "Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility." *Science* 348 (6242):1422-5. doi: 10.1126/science.aab2374.

Okasha, Samir. 2016. "On the interpretation of decision theory." *Economics & Philosophy* 32 (3):409-33.

Ormans, Laurent. 2016. "50 Journals used in FT research." accessed September 1. https://www.ft.com/content/3405a512-5cbb-11e1-8f1f-00144feabdc0.

Ridgeway, Cecilia L., and Shelley J. Correll. 2006. "Consensus and the creation and status beliefs." *Social Forces* 85 (1):431-53.

Rubin, Hannah, and Cailin O'Connor. 2018. "Discrimination and Collaboration in Science." *Philosophy of Science* 85:380-402.

San Francisco Declaration on Research Assessment. 2013. "The San Francisco Declaration on Research Assessment (DORA)." accessed September 1. https://sfdora.org/read/.

Sauder, Michael, and Wendy Nelson Espeland. 2006. "Strength in numbers? The advantages of multiple rankings." *Indiana Law Journal* 81 (1):205-27.

Schimmack, Ulrich. 2015. "Replicability Ranking of 26 Psychology Journals." January 18. https://replicationindex.wordpress.com/2015/08/13/replicability-ranking-of-26-psychology-journals/.

Science. "Mission and Scope." accessed September 1. http://sciencemag.org/about/mission-and-scope.

Strevens, Michael. 2003. "The role of the priority rule in science." *Journal of Philosophy* 100 (2):55-79.

West, Jevin D., Theodore C. Bergstrom, and Carl T. Bergstrom. 2010. "The Eigenfactor Metrics™: A network approach to assessing scholarly journals." *College & Research Libraries* 71 (3):236-44.

Willer, Robb, Ko Kuwabara, and Michael W. Macy. 2009. "The False Enforcement of Unpopular Norms." *American Journal of Sociology* 115 (2):451-90.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. 2015. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.

Wilsdon, James, Judit Bar-Ilan, Robert Frodeman, Elisabeth Lex, Isabella Peters, and Paul Wouters. 2017. Next-generation metrics: Responsible metrics and evaluation for open

science. Report of the European Commission Expert Group on Altmetrics. European Commission.

Zollman, Kevin J. S. 2018. "The Credit Economy and the Economic Rationality of Science." *The Journal of Philosophy* 115:5-33.

Zuckerman, Harriet, and Robert K. Merton. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System." *Minerva* 9 (1):66-100.

Function Words and Context Variability*

Shane Steinert-Threlkeld
S.N.M.Steinert-Threlkeld@uva.nl

Draft of 30 October 2018. Comments Welcome!

> *Twas* brillig, *and the* slithy toves
> *Did* gyre *and* gimble *in the* wabe;
> *All* mimsy *were the* borogoves,
> *And the* mome raths outgrabe.
>
> Excerpt from 'Jabberwocky', in Carroll (1871), emphasis mine.

The poem excerpted in the epigraph has often been called a 'nonsense poem'. After all: what does it mean? What is a slithy tove? What does it mean to be brillig or mimsy? Calling it nonsense, however, overlooks the amount of meaning we can extract from the emphasized words: minimally, a scene in the past is being described, which took place somewhere called a 'wabe'. The emphasized words are what are known as *function words*: they provide the 'grammatical glue' among the *content words*, which are indeed nonsense in this excerpt.

The distinction between these two types of expression occupies a central place in modern theoretical linguistics. Rightfully so: every natural language exhibits a distinction between function and content words. Yet surprisingly little has been said about the emergence of this universal architectural feature of natural languages. Why have human languages evolved to exhibit this division of labor between content and function words? How could such a distinction have emerged in the first place?

This paper takes steps towards answering these questions by presenting a simple model of trial-and-error language learning in which a division of signals into function and content words emerges. In the next section, I briefly but more explicitly introduce the distinction. In Section 2, I argue that a necessary condition for the emergence of the distinction is the presence of *non-trivial composition* (in a sense to be made precise). I present three case studies in which only trivial composition emerges and a mathematical result that diagnoses why that is the case. In Section 3, I introduce a new type of signaling game – the Extremity Game – in which the objects of communication vary from play to play. Amidst such variation, a distinction between function and content words could be useful. Section 4 reports an

*Acknowledgments to be added. This work was supported by funding from the European Research Council under the European Unions Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

experiment, in which artificial neural networks are trained by reinforcement learning to communicate in the Extremity Game. The emerging languages are analyzed: when the agents can pay attention to perceptually salient features of the context, they learn a system with complex signals that we can interpret as a gradable adjective plus a superlative morpheme (a prime example of a functional item). Section 5 concludes.

1 Functional and Lexical Categories

Modern theoretical syntax distinguishes between two broad types of syntactic categories: lexical and functional.[1] The former broadly correspond to the major parts of speech: nouns, verbs, adjectives, and adverbs. The latter are a bit more varied, but include:[2]

- Prepositions: 'in', 'above', 'from', 'to', ...
- Determiners: 'a', 'the', 'every', 'some', 'many', ...
- Conjunctions: 'and', 'or', ...
- Complementizers: 'that', 'if', 'whether', ...
- Tense:
  - Auxiliaries: 'have', 'is', 'was', ...
  - Modals: 'will', 'would', 'can', 'might', 'ought', ...

Exactly characterizing the distinction remains tricky. After observing that lexical categories have 'contentful' meaning, while functional categories have 'grammatical' meaning,[3] it is usually observed that the former constitute *open classes* and the latter *closed classes*. Roughly: one can very readily introduce new nouns and verbs to a language as needed. By contrast, trying to introduce a new preposition 'belove' meaning partially above and partially below would be quite difficult. Kaplan (1978) famously tried to introduce a new expression 'dthat', which rigidly referred to the satisfier of a description. Though he ably demonstrated the use of such a tool, that it never caught on can be partially attributed to the fact that demonstratives belong to a closed class. For the purposes of this paper, these distinctions suffice to point to the intended contrast.

Before proceeding, it's worth highlighting that the field of semantics – the scientific study of linguistic meaning – roughly divides itself along the lexical/functional line as well. The tradition descending from Montague via Partee and many others, usually called formal semantics, studies specifically compositional semantics. A survey of the textbooks in this field[4] shows that the major expressions studied come exactly from the functional categories.

[1] See, for example, pp. 43-46 of the textbook Carnie (2006).

[2] This list is incomplete and meant to be illustrative only. There are some debates about exactly which category certain expressions belong to, but they are orthogonal to present concerns.

[3] See, e.g., Carnie (2006) and Rizzi and Cinque (2016). Muysken (2008) is a thorough overview of functional categories.

[4] For example, Heim and Kratzer (1998) and Jacobson (2014).

Lexical semantics – the study of the meanings of basic expressions – studies at length the meanings of individual expressions and groups thereof in the lexical categories.[5] Seen in this light, explaining the emergence of the distinction between functional and lexical categories occupies a central role in the broader explanation of the emergence of compositionality.

2 Non-Trivial Composition

In this section, I build on the foregoing remarks in order to argue for the following claim: for a communication system to have function words, there must exist *non-trivial composition* (in a sense to be made precise) of complex signals. After presenting this argument, I will analyze three case studies from the literature on the evolution of compositionality which exhibit only trivial composition. The reasons for this are then made precise in the form of a triviality result: given the assumptions about optimal communication often made, the resulting systems must be trivially compositional.

The principle of compositionality says that the meaning of a complex expression is determined by the meanings of the parts and how they are put together.[6] Natural languages are compositional: whence the ability of competent speakers to produce and comprehend a potentially infinite set of novel expressions. A language can, however, be compositional without exhibiting the rich flexibility that human languages do. We will use the following definition:[7]

(1) A communication system is *trivially compositional* just in case complex expressions are always interpreted by intersection (generalized conjunction) of the meanings of the parts of the expression.

The force of this definition can be brought out by an example: Titi monkey calls.[8] In a series of predator-model experiments, it was found that raptors in the canopy elicit sequences of A calls, cats on the ground elicit sequences of B calls, cats in the canopy elicit one A followed by a sequence of Bs, and raptors in the canopy elicit a sequence of As followed by a sequence of Bs. While the full details do not concern us,[9] Schlenker, Chemla, Schel, et al. (2016a) argue that the best analysis of this call system involves the following semantics, interacting with some plausible pragmatic principles:

(2) Compositional semantics of Titi alarm calls: where t is a time,
   a. $[\![B]\!]^t = 1$ iff there is a noteworthy event at t
   b. $[\![A]\!]^t = 1$ iff there is a serious non-ground alert at t
   c. $[\![wS]\!]^t = 1$ iff $[\![w]\!]^t = 1$ and $[\![S]\!]^{t+1} = 1$
   [where w is a call and S a sequence of calls]

The crucial feature of this semantics concerns the rule (2c) for interpreting complex expressions (sequences of calls). It says that a sequence of calls is interpreted by first evaluating

[5] See, for example, Levin and Rappaport Hovav (2005).
[6] Frege (1923), Janssen (1997), Pagin and Westerståhl (2010a), and Pagin and Westerståhl (2010b).
[7] For this use, see Schlenker, Chemla, Schel, et al. (2016b) and Zuberbühler (2018).
[8] Cäsar et al. (2013) and Schlenker, Chemla, Schel, et al. (2016a).
[9] See Steinert-Threlkeld (2016b) for some reservations about the full analysis.

the beginning of the sequence at time t, then evaluating the rest of the sequence at time $t+1$, and conjoining the results. This clause results in the following: each call in the sequence contributes to the meaning of the whole *independently* of the other calls, with the complete meaning resulting from conjunction. It thus constitutes a paradigm of the definition of trivial compositionality in (1).[10]

In other words, non-trivial compositionality involves non-conjunctive modification of one linguistic item by another. Examples of such systems can also be found in communication systems much simpler than human language. In particular, Campbell's monkeys have been argued to exhibit it.[11] They have two basic alarm calls: an eagle call *hok* and a general alert *krak*.[12] Moreover, both calls combine with what appears to be a suffix *-oo*, which has the effect of weakening the severity of the calls. Schlenker, Chemla, Schel, et al. (2016a) propose the following semantics:

(3) $[\![R\text{-}oo]\!]^t = 1$ iff at t the sender is alert to a disturbance that licenses R but that is not strong among such disturbances.

This is non-trivial: *-oo* does not contribute independent meaning that is then conjoined with the contribution of *hok* or *krak*. Rather, it combines with one of the latter calls to modify the normal meaning of that call.

Here is the simple argument for the claim that non-trivial composition is necessary for the emergence of function words. Recall the characterization thereof as 'grammatical glue': they precisely do not contribute independent content to a sentence, but structure that provided by the content words. In a trivially compositional communication system, each expression contributes independent meaning to the complex expressions containing it. Therefore, none of the expressions therein are function words.

Before proceeding, we note that the presence of non-trivial composition does not suffice for the presence of function words. To see this, consider subsective adjectives.[13] These are adjectives like 'skillful', which have the property that for every noun, a 'skillful N' is an N, but is not 'skillful' in any sense independent from the noun. For example:

(4) a. Jakub is a skillful rock climber.
b. Jakub is a cook.
c. Therefore, Jakub is a skillful cook.

The inference pattern in (4) is not valid: Jakub can be skillful at one thing but not at another. If 'skillful' contributed its meaning independently of the noun it combines with, the inference would be valid: Jakub would be a climber, a cook, and skillful; therefore, a skillful cook. But 'skillful' is still a content word. One could imagine a very simple language whose only complex expressions were of the form 'Adj N', but which had subsective adjectives. This language would be non-trivially compositional but would have no function words.

[10] Berthet et al. (2018) argue that the proper semantics for Titi calls is not in fact trivially compositional. Nevertheless, the presentation just given illustrates what such a system would look like.

[11] Ouattara, Lemasson, and Zuberbühler (2009) and Schlenker, Chemla, Arnold, et al. (2014).

[12] The possibly different meaning of *krak* in different habitats of Campbell's monkeys is the subject of the aforementioned papers. We follow Schlenker, Chemla, Schel, et al. (2016a) in giving it a general meaning.

[13] Partee (1995).

Now, I will present three case studies of prominent models purporting to explain aspects of the evolution of compositional communication. Each of them, however, will turn out to exhibit only trivial composition. After presenting the case studies, I identify common underlying assumptions and then prove a mathematical fact demonstrating that under those assumptions, the resulting communication systems must be trivially compositional. In light of the foregoing, none of these extant approaches can explain the emergence of the distinction between function and content words.

2.1 Three Études

Nowak and Krakauer (1999) apply mathematical models of natural selection to the evolution of language, providing conditions under which a 'grammatical' language will evolve from a non-compositional one. In their model, states are object-action pairs, loosely modeling events. They compare two types of languages: one in which each object-action pair has an independent label, and another in which each object has a corresponding expression, each action has a corresponding expression, and the agents communicate by sending the corresponding pair of expressions to communicate about an object-action pair. While the results they obtain are indeed interesting, it should be clear from this brief exposition that the type of language that they consider exhibits only trivial composition: each component of a complex expression contributes its bit of meaning (either an object or an action) independently of the other.

Barrett (2007) and Barrett (2009) studies a generalization of signaling games[14] with multiple senders. In the simplest case, there are four states of nature and two sender, each of whom can send one of two signals to one receiver. The senders, but not the receiver, know which state obtains. Simulations show that a simple form of reinforcement learning leads these agents to a situation of perfect communication. Given the nature of the setup, the resulting systems look as follows. One sender partitions the four states into two sets of two, one for each signal. The other sender sends its two signals in an *orthogonal* partition.[15] One can imagine the states as a two-by-two square, with one sender indicating the row and the other the column of the true state. Such a system again exhibits only trivial composition, since the meaning of each sender's signal is independent of the other's and the receiver interprets the sequence by intersecting the two.

Finally, Mordatch and Abbeel (2018) study the emergence of communication in a multi-agent setting where each agent has a private goal that it wants to achieve.[16] The agents – which are in this case recurrent neural networks – communicate about a world with various colored landmarks in it. Each agent additionally has a color and its own perspective from its position (i.e. no agents share a frame of reference). The goals consist of getting an agent to perform an action (going to or looking at) at one of the landmarks. With appropriate costs for maintaining large lexicons, the agents learn to send sequences of signals with separate signals for which agent, which action, and which landmark. These three types of signals have independent meanings, which are combined by conjunction.

[14] Lewis (1969) and Skyrms (2010).
[15] See, e.g., Lewis (1988).
[16] The set of goals is assumed to be consistent, i.e. all of the goals are simultaneously realizable.

2.2 A Limitative Result

There is in fact an underlying reason that these systems exhibit only trivial composition. Although the three cases just illustrated come from different theoretical frameworks, they all share the same following assumptions:

(A1) Agents communicate about a fixed set of states. (Object/action pairs, separate points of a state space, and agent/landmark/action tuples, respectively.)

(A2) Optimal communication consists in correctly identifying the true member of the state space.

(A3) Messages are fixed-length sequences of signals from fixed sets.

It turns out that under these assumptions, there's a mathematical sense in which optimal communication will be trivially compositional. This is captured in the following result:

(5) Let X and $\{M_i\}_{i\in I}$ be any sets, and f, g two functions of the following type:

$$X \xrightarrow{f} \prod_i M_i \xrightarrow{g} X$$

Define $f_i^{-1}(\overrightarrow{m}) := \{x \in X : f(x)_i = \overrightarrow{m}_i\}$. Then the following holds.

$$\text{If } g \circ f = \mathrm{id}_X, \text{ then for all } \overrightarrow{m}, \{g(\overrightarrow{m})\} = \bigcap_i f_i^{-1}(\overrightarrow{m})$$ [17]

Here, X represents the fixed set of states about which the agents communicate. Note that the structure of this set does not matter. $\prod_i M_i$ is the set of possible sequences of signals, with each M_i being the signals available to be sent in position i of a sequence. f is a sender function: a function from states to sequences of signals. This can capture a single sender, or multiple acting either independently or in concert. g is a receiver function: it decodes the sequence of signals to one of the states X. Because id_X is the identity function on X, mapping each point to itself, that $g \circ f = \mathrm{id}_X$ means that optimal communication has been achieved, in the sense that the receiver always recovers the true state from X. Under that assumption, the result says that the receiver interprets a complex message (a sequence) by *intersecting* the independent meanings of each signal in the sequence (represented by $f_i^{-1}(\overrightarrow{m})$).

This result identifies three assumptions that cannot all be maintained if one wants to model the emergence of non-trivial composition, which I have just argued is a necessary step for explaining the emergence of function words. Not every approach makes all three of these assumptions. In particular, Steinert-Threlkeld (2014) and Steinert-Threlkeld (2016a) as well as Barrett, Skyrms, and Cochran (2018) drop (A3). In these models, not every message is a

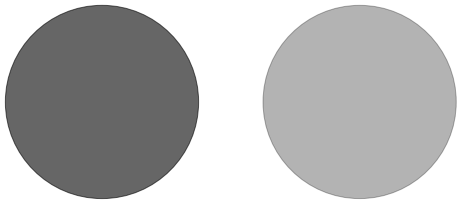
[17] *Proof:* Note first that g must be a surjection and f an injection. Without the former, there would be an $x \in X$ that is not $g(\overrightarrow{m})$ for any $\overrightarrow{m}$, and so $g \circ f \neq \mathrm{id}_X$. Without the latter, distinct points in X would get mapped to the same point in X by $g \circ f$. Now, suppose there were an $\overrightarrow{m}$ such that $\{g(\overrightarrow{m})\} \neq \bigcap_i f_i^{-1}(\overrightarrow{m})$. This can hold only if $\bigcap_i f_i^{-1}(\overrightarrow{m})$ contains more than one element, since $g(\overrightarrow{m})$ has to belong to the intersection. This entails that there is another point $x \neq g(\overrightarrow{m})$ for which $f(x) = \overrightarrow{m}$, contradicting the injectivity of f. □

sequence of the same length. In the former, one sender can choose whether or not to prefix a set of signals with an additional signal. In the latter, two senders choose *whether or not* to send a signal, so messages can be either of length one or two. In either case, the message space is a union, not a product (i.e. not of the form $\prod_i M_i$ for any sets M_i), and so the limitative result does not apply.

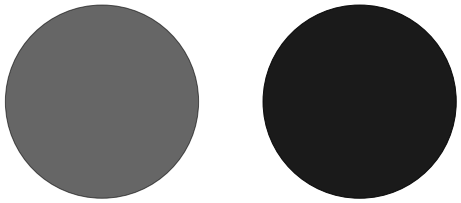
In the remainder, I will develop a model which maintains (A2) and (A3) but drops the assumption (A1) of a *fixed* set of states that the agents communicate about. That is: the context in which the agents are communicating will vary. Against that backdrop, there will be a role for function words to play.

3 A Signaling Game with Varying Contexts

The variant on the signaling game that I will use to illustrate the emergence of function words will have the agents talking about varying sets of objects with multiple *gradable properties*. To get a feel for the kind of task involved, consider the following adaptation of an example from Graff (2000).[18] Suppose that we are both looking at the following two circles, drawn on top of a table.



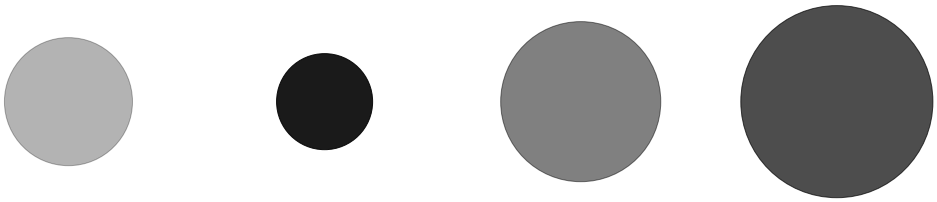
For whatever reason, you need me to put something on the left circle. You might say "put it on the *darker circle*". By contrast, suppose that you had the same communicative needs, but now the circles on the table looked as follows.



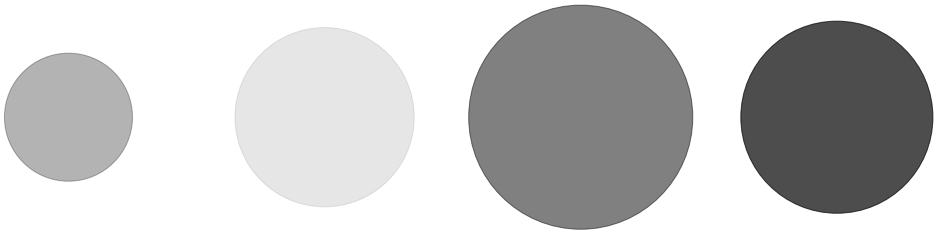
Now, to tell me to put it on the left circle, you might say "put it on the *lighter circle*", or a bit more circuitously "put it on the less dark one".

The target referent of your communication – the circle on the left – has exactly the same size and shade in both contexts. But in one context, it's dark*er* than the other circle, while in the other context, it's light*er*. (For the purposes of illustration, we can assume that you could not refer to the circles by their spatial position or demonstratively. If you'd like: your friend is looking at a picture on a screen that may have been scrambled.) Finally, we can imagine similar situations with more than one gradable property. Suppose, again, that you need to communicate about the leftmost circle in the following array.

[18] See Syrett, Kennedy, and Lidz (2010) for a study using similar contexts with children.



Here, it's natural to call the leftmost circle "the lightest one". Now consider the following context.



Here, you are likely to refer to the circle on the left as "the smallest one". These situations have the following structure: in each context, each object has two very salient gradable properties: a size (radius) and a darkness. These dimensions distinguish the target object: it has either the largest or the smallest value in one of those dimensions. By drawing attention to that fact, one can successfully refer to it. Moreover, you can do so in a very economical way: with labels for the properties and morphemes like the superlative *-est* (and its corresponding negative counterpart, 'the least'), successful communication is ensured. This is done without talking about specific degrees of size or of lightness and in a way in which an object with exactly the same degrees on all relevant properties will be referred to in different ways in different contexts.

I will convert communicative scenarios like the above into a type of signaling game – called the Extremity Game – with a few helper definitions. Following the literature on gradable adjectives,[19] I will assume that objects have some number of gradable properties, where each property has a corresponding *scale*. A scale in turn is a set of *degrees*, totally ordered with respect to a dimension. For example, the size of a circle corresponds to its radius, with degrees being positive real numbers (i.e. $\mathbb{R}^+$). For the degree of an object o on a scale s, I will write $s(o)$. Given a set S of scales, I will define a context as follows.

(6) A *context* c over scales S is a set of objects such that: for each $o \in c$, there is a scale $s \in S$ such that either o has the least degree on s ($o = \arg\min_{o' \in c} s(o')$) or the highest degree on s ($o = \arg\max_{o' \in c} s(o')$).

At its most general form, the game takes place between a sender and a receiver in the following way.

(7) Extremity Game, in general:
   a. Nature chooses a context c and a target object $o \in c$.
   b. The sender sees c and o and sends a message m from some set of messages M.

[19] See, for instance, Kennedy and McNally (2005) and Kennedy (2007) and the references therein.

c. The receiver sees c and m and chooses an object o' from c.

d. The play is successful (and the two agents equally rewarded) if and only if $o' = o$.

To fully specify a game, one must say what the messages M available are and how the agents make their choices. I will specify the former now and the latter in the next section. The set of available messages will be inspired by the semantics for gradable adjectives. There, it is assumed that adjectives map objects (of type e) on to their degree on the corresponding scale (of type d). Morphemes like *-est* and *least* then map a contextually specified set of objects to the subset with the highest and lowest degrees.

(8) Toy semantics for a gradable adjective and superlative morphemes.

a. $[\![\text{size}]\!] = \lambda x.s_{\text{size}}(x)$

b. $[\![\text{-est}]\!]^c = \lambda P_{\langle e,d\rangle}.\lambda x_e.x \in c \text{ and } \forall x' \in c, P(x) \succeq P(x')$

c. $[\![\text{least}]\!]^c = \lambda P_{\langle e,d\rangle}.\lambda x_e.x \in c \text{ and } \forall x' \in c, P(x) \preceq P(x')$

Now, for the crucial observation: in contexts as defined in (6), having one expression for each scale and the morphemes *-est* and *least* will suffice to uniquely pick out each object in the context. I will assume, then, that the set of messages $M = M_S \times M_P$ where M_S is a set of size $|S|$ (i.e. there are as many messages in M_S as there are gradable properties for each object) and M_P is a set of size two (P for 'polarity'). The players of an Extremity Game will be able to successfully communicate if they can learn to associate each message in M_S with a distinct scale and the two signals in M_P with something akin to *-est* and *least*. As advertised, this setup meets two of the three assumptions in the limitative result (5) – (A2) optimal communication is correct identification of a target object and (A3) messages come from a product space – but drops (A1): because the context varies from play to play of the game, there is no fixed set of objects about which the agents communicate.[20]

4 Experiment

The goal is to show how a simple semantic system like 8 could emerge via a simple dynamics among agents playing an Extremity Game. In particular, we will use *reinforcement learning*:[21] agents make choices, receive some reward (in our case, for successful communication of the target object in context), and adjust their behavior so that they are more likely to make the corresponding choices in the future.

While most approaches to reinforcement learning in signaling games use a variant of a simple algorithm called Roth-Erev learning,[22] such an algorithm will not suffice for present purposes. On this approach, choices are reinforced entirely independently of one another. Two factors of the present setup require a stronger method. On the practical side, there is a combinatorial explosion that comes from having variable contexts with multiple objects that have multiple gradable properties: there are so many contexts that most of them will not be seen often enough for such an algorithm to be effective. On the conceptual side, if

[20] While the agents in an intuitive sense communicate 'about' a fixed set of objects – all objects with $|S|$ gradable properties – each communicative exchange concerns a different subset thereof.

[21] Sutton and Barto (2018)

[22] Roth and Erev (1995)

choices are reinforced entirely independently, there will be no pressure for signals to emerge that group objects based on the degrees of various properties and their relative position on scales in context.

To overcome this limitation, I will use a type of agent with a built-in capacity for stimulus generalization: artificial neural networks.[23] This choice was made because such networks provide a simple, widely used, and somewhat biologically plausible model that has the capacity to generalize. Other approaches to stimulus generalization in learning in signaling games use a method called *spill-over*.[24] In that framework, not only are the actual choices reinforced, but so too are *similar* choices in similar choice points. Exactly how reinforcement works thus depends on definitions of similarity between choices and between states. While some domains provide natural such definitions,[25] it is not immediately obvious how to define how similar one context-target pair is to another in an Extremity Game. Neural networks will learn to treat certain pairs as similar and others not, without the theorist having to hard-wire a definition of similarity into the learning model.[26]

4.1 Methods

A trial of our experiment will consist of some number of iterations of playing an Extremity Game as in (7). The sender and receiver are each neural networks, schematically depicted in Figure 1. They are trained using the REINFORCE algorithm, the simplest in a family of methods known as policy gradient methods.[27] The intuition behind this algorithm is just as before. Consider the sender. The sender is a policy that takes as input a context and a target and outputs a probability distribution over messages (in this case, two distributions: one over M_S and one over M_P). The sender's policy is parameterized by the weights and biases that connect the neurons in the network. Thanks to what is known as the policy gradient theorem, modern variants of stochastic gradient descent can be used to adjust the weights and biases in a way that is guaranteed to make positively reinforced actions sampled from the policy more likely in the future.

We varied the number of dimensions (i.e. gradable properties) between 1 and 3, and ran 10 trials for each. We trained for five-, twenty-, and fifty-thousand mini-batches respectively, where each mini-batch was size 64. In other words, the agents play 64 games in between each update of their policies; this reduces the variance in learning. We also experimented with two different neural architectures for the receiver – called Basic and Attentional – for reasons that will become clear in what follows. We recorded the rolling accuracy over 10 training steps, as well as the accuracy and detailed properties about contexts and signals used on 5000 new games at the end of training.

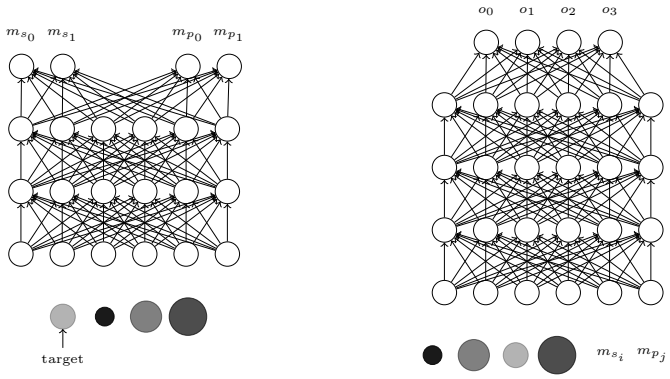
[23] Nielsen (2015) and Goodfellow, Bengio, and Courville (2016)

[24] See O'Connor (2014). The name 'spill-over' comes from Franke (2016).

[25] For instance, if the goal is to choose a point on a line, the distance between the true point and the guessed point is very natural.

[26] See Lazaridou, Peysakhovich, and Baroni (2017) for a similar approach, which inspired the present one. Their contexts consist of two natural images, one of which is the target. The sender chooses one signal from a fixed-sized vocabulary to send to the receiver. While they are interested in whether natural concepts emerge in such a setting, I am focused on less natural input but more complex communication structures in order to explore the emergence of functional vocabulary.

[27] Williams (1992). See chapter 13 of Sutton and Barto (2018) for a modern introduction.



(a) Basic Sender (b) Basic Receiver

Figure 1: Schematic depictions of basic network architectures. The input is on the bottom, followed by a sequence of hidden layers to output layers on the top. The output neurons produce probabilities of choosing the action written above them.

Complete details of the network architectures and training set-up are included in an Appendix. The code and data can be found at https://github.com/shanest/function-words-context.

4.2 Results: Basic Receiver

The learning curves over training for each trial of each dimension, with the Basic Receiver, are plotted in Figure 2. As can be seen, in the one- and two-property cases, the agents learn to communicate nearly perfectly in a relatively short amount of training steps. By contrast, in the three-dimension case, the agents do not regularly achieve a high degree of communicative success after 50000 mini-batches. The mean success rates on 5000 new games at the end of training time are reported in Table 1.

In the one-dimensional case, the context consists of two objects that have a property to different degrees. The successful communication protocol that the agents learn to use reliably sends one signal when the target has the lower degree and the other signal when the target has the higher degree.

dims	mean	std
1	0.975	0.006
2	0.985	0.003
3	0.731	0.062

Table 1: Accuracies on novel games.

In the two-dimensional case, things are not quite as aligned with expectations. Figure 3 shows a typical communication protocol that emerges in the two dimensional case. The colored bars correspond to the particular signals sent. The left column corresponds to M_S and the right column to M_PThe colored bars correspond to the particular signals sent. The left column corresponds to M_S and the right column to M_P. In the top row, the x-axis corresponds to the 'true' dimension of the target object (i.e. the dimension for which the target had an extreme value in context). In the bottom row, the x-axis corresponds to the 'true' polarity of the target object (i.e. whether it had the true property to the least or highest degree).

The bottom-left cell shows an interesting pattern: the message from M_S sent always

Figure 2: Learning curves for basic sender and receiver.

corresponds to the true polarity (minimum or maximum). This is because one message is always sent when the true polarity is 0 (minimum) and the other when the polarity is 1 (maximum). Unfortunately, that the top row shows no such separation implies that no signal is being used to communicate the 'true' dimension. The equal heights of all the bars in the top row imply that the two messages in M_S (left column) and in M_P (right column) are used an equal number of times when the true dimension is 1 and when the true dimension is 0.

In fact, closer inspection reveals the following: the learned communication systems are always 'maximally' separating in the following sense: for any two contexts c, c' and targets o, o', if $o = \arg\min_c s_d(o)$ and $o' = \arg\max_{c'} s_d(o)$ for the same dimension d, then the sender's message for o in c differs from its message for o' in c' in both syntactic positions. This holds true for the 3-dimensional case as well. Figure 4 shows an example learned system. The bottom-right cell shows that the agents do use M_P to distinguish the true direction of the target. But the top-left cell shows that the agents do not associate different signals in M_S with different dimensions: rather, they separate targets in the way just described.

These results show that basic senders and receivers do not, under the REINFORCE algorithm, learn to communicate in accord with the toy semantics in (8). One might think that one of the messages still looks like a superlative morpheme, since it reliably correlates

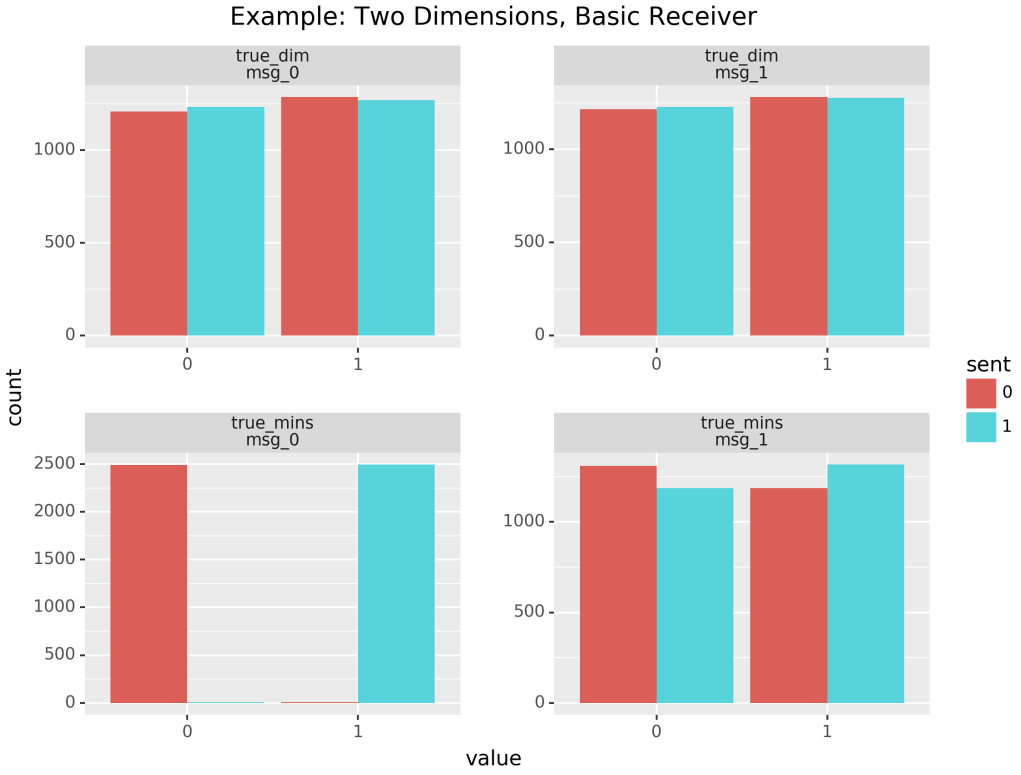


Figure 3: Example communication system with basic receiver and two dimensions.

with the true direction of the target object. While this is indeed very interesting and does show that the networks are clustering objects on the basis of their direction (for example, they never separate on dimension and group together based on direction), given that they do not use the other signal to communicate the true dimension, it does not look like there's non-trivial modification of one linguistic item by another.

4.3 Results: Attentional Receiver

Intuitively, the networks are not learning to use a signal to group objects together based on dimension. This could be for roughly the following reason: in expectation, target objects that differ only in whether they are the minimum/maximum in context on a dimension will actually be farther from each other in Euclidean space than from other objects. Because of this, it could be that the agents use maximally different signals for the two types of target objects.

To help the agents learn to communicate based on the dimension, I will use what is known as an *attention mechanism* in machine learning.[28] Intuitively, a neural network can

[28]See, for instance, Mnih et al. (2014) and Xu et al. (2015).

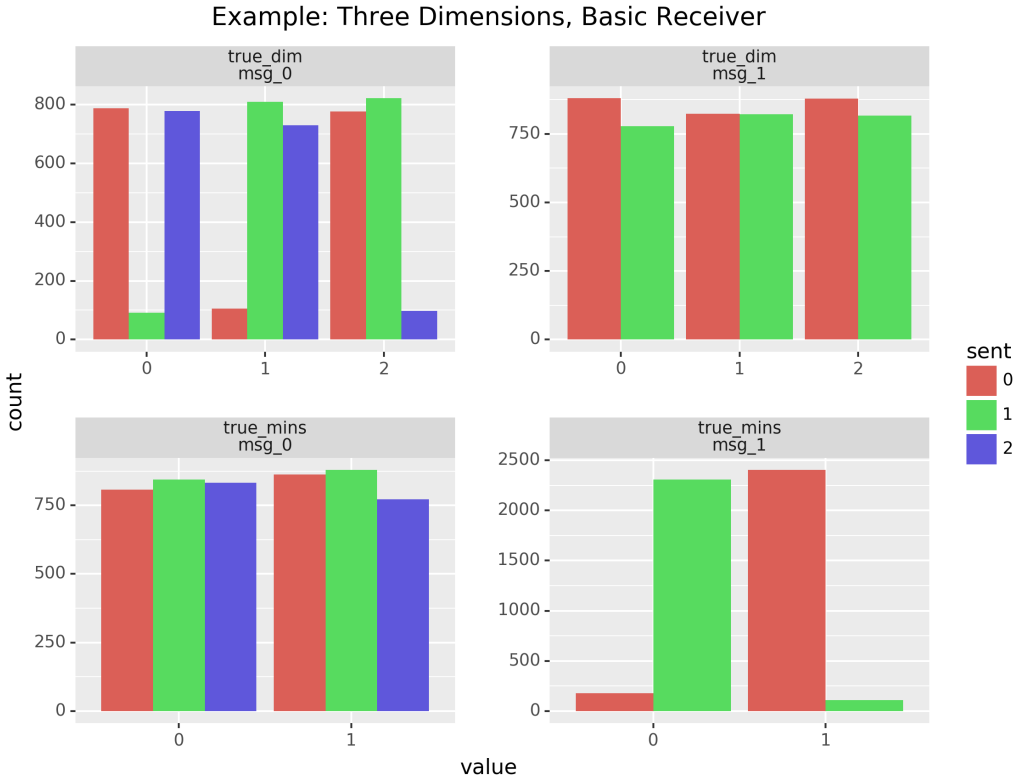


Figure 4: Example communication system with basic receiver and three dimensions.

learn to pay more or less attention to different portions of its input. The network (or a sub-component thereof) computes a weighting of the input positions which is then used to filter the actual input. The weight can be 'hard' – selecting a sub-region of the input – or 'soft' – re-weighting the input so that different nodes are more or less attended to than in the raw input.

One can think of attention as reflecting something like perceptual salience: the network can learn to focus its attention on salient features of its input, since those features are likely to help it solve its task. For instance, a neural image caption generator with attention will likely focus its attention on well-defined objects in an input image. These salient objects are likely to help it generate a plausible caption.

Attentional Receivers, as I will develop them, implement a hard attention mechanism in the following sense. First, they receive as input the context c and the message m_{s_i} from M_S chosen by the sender. On this basis, the receiver *chooses a dimension to attend to*: the input is filtered so that the agent only sees the objects according to one dimension (e.g. size or lightness). Then, the agent uses this attended-to dimension and the message from M_P chosen by the sender to choose a target object. This attention mechanism reflects the perceptual salience of the gradable properties of the objects: it is very natural, for instance,

in the contexts in Section 3, to attend only to the size or the shade of the circles. Figure 5 depicts this architecture.

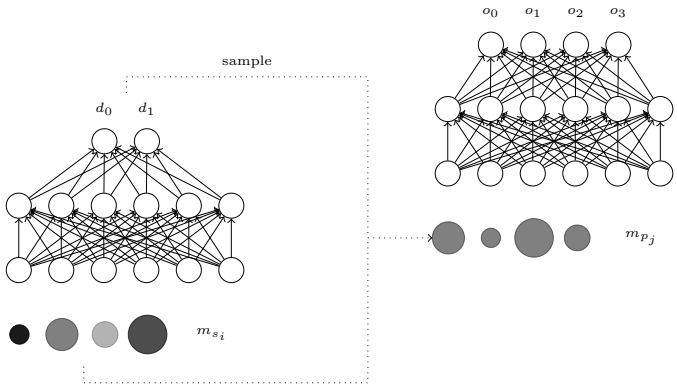


Figure 5: Attentional Receiver architecture, schematically. The receiver first chooses a dimension to attend to, then chooses a target based only on that dimension. In this schematic, the chosen dimension is size; differences in shading have been washed out by the attention mechanism.

The learning curves over training for each trial of each dimension – but with a Basic Sender and Attentional Receiver – are plotted in Figure 6. The mean success rates on 5000 new games at the end of training time are reported in Table 2. As before, in the one- and two-property cases, the agents learn to communicate nearly perfectly in a relatively short amount of training steps. In all cases, it appears that learning is a bit slower than with basic receivers. This makes perfect sense: an attentional receiver has to learn two types of choices to make, as opposed to just one. In the three dimensional case, the attentional receiver achieves a high-degree of accuracy more frequently than the basic receiver, but also gets stuck in sub-optimal states more frequently.

The resulting communication protocols behave exactly like the toy semantics in (8). Figure 7 shows an example protocol in two dimensions. Here, the top-left cell shows that the choice of signal from M_S reliably communicates the true dimension: when the dimension is 0, the sender chooses m_{s_0} and when the dimension is 1, the sender chooses m_{s_1}. Similarly, the bottom-right cell shows that the choice of signal from M_P signal reliably communicates the true direction (i.e. whether the target has the relevant property to the largest or smallest degree). Figure 8 shows an example learned communication system in three dimensions. Again, in complex signals, one signal communicates a dimension, and the other communicates whether the target has the most or least degree on the corresponding scale.

dims	mean	std
1	0.959	0.005
2	0.964	0.005
3	0.697	0.144

Table 2: Accuracies on novel games.

When the agents are communicating in this way, the signals that communicate direction can be interpreted as function words. The signals in M_S reliably communicate a bit of 'content': a dimension. The signals in M_P reliably signal whether the target has the greatest/lowest degree *along that dimension* of all the objects in the context. This is non-trivial

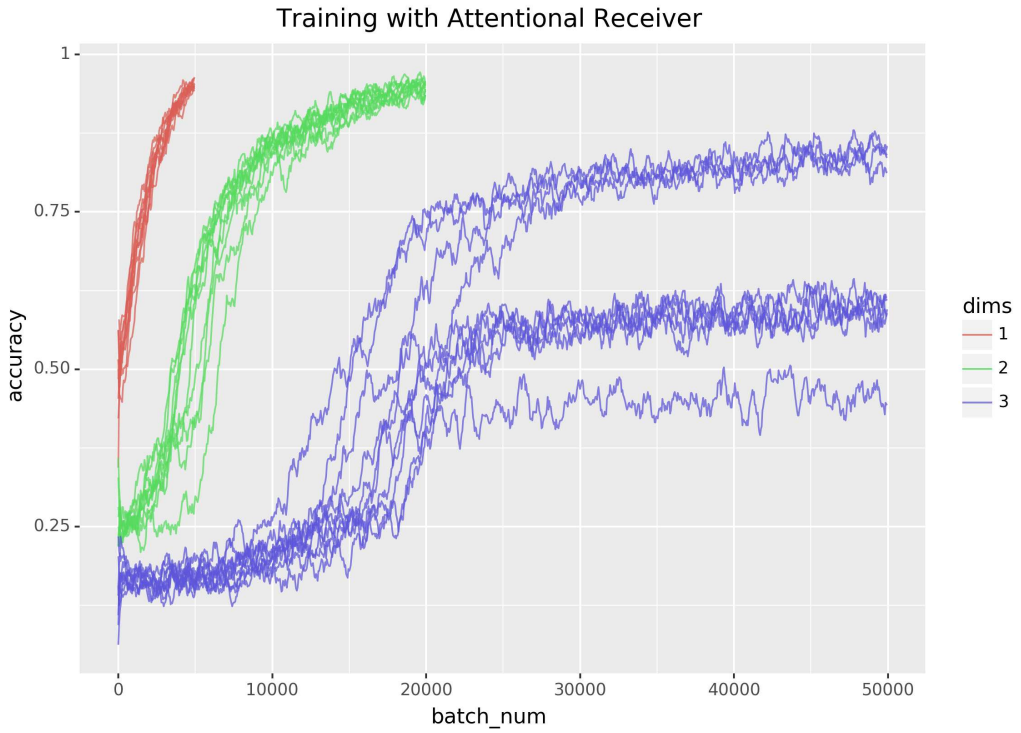


Figure 6: Learning curves for basic sender and attentional receiver.

modification of one linguistic item by another. Thus, when the receiver knows to use one of the signals to attend to a particular dimension in context, the two agents can learn to use their signals in a non-trivially compositional way.

5 Conclusion

Let us take stock. After introducing the distinction between functional and lexical categories, I argued that there are in principle reasons why many extant models of the evolution of compositionality cannot explain the emergence of function words: given their assumptions, they can only explain trivial composition; but non-trivial composition is a necessary precondition for the presence of function words. I then introduced a signaling game with variable contexts consisting of multiple objects with varying gradable properties. Simple reinforcement learning by neural networks – in particular with the ability to pay attention to certain perceptually salient aspects of the input – in this game can generate expressions that are appropriately characterized as function and as content words.

Much work remains to be done. One would like neural architectures that make fewer assumptions about what aspects of the input the receiver pays attention to. A first step in this direction will be to use a soft, as opposed to hard, attention mechanism. A more

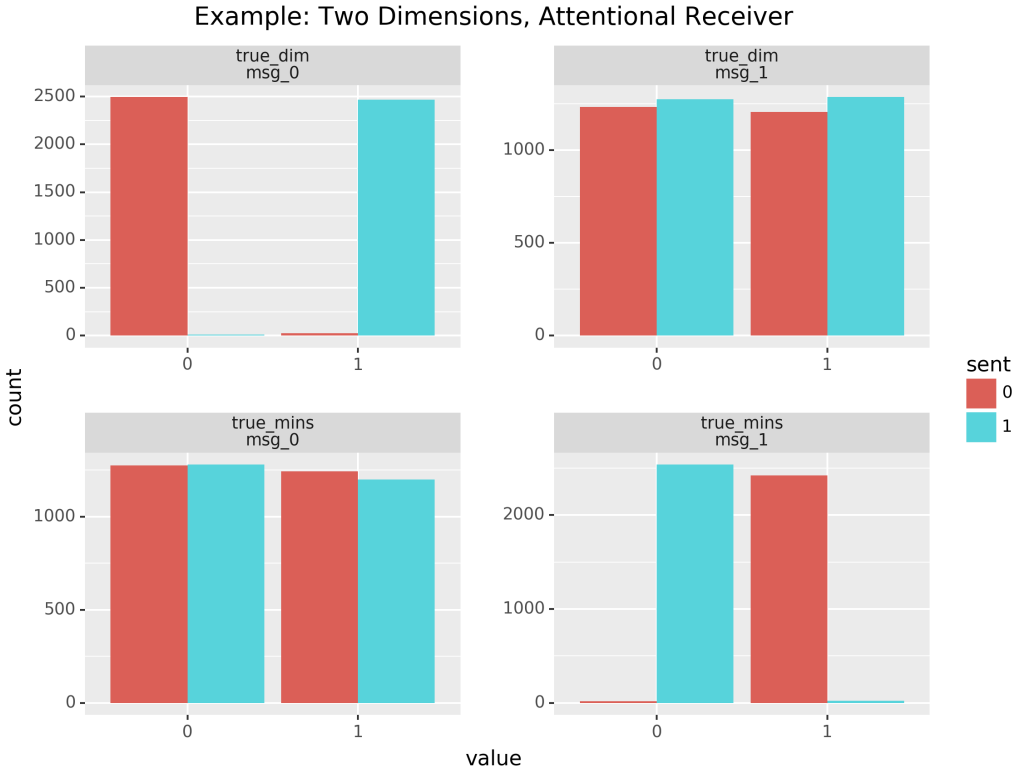


Figure 7: Example communication system with attentional receiver and two dimensions.

thorough hyper-parameter search may also generate more reliable learning results in the higher-dimensional setting. One can also generalize the input so that the networks also have to discover *which dimensions* are relevant for being able to successfully refer to objects across contexts, instead of having it built into the current definition of context. More generally, one would like communication systems like those exhibited here to emerge in the very general setting of communicating by a sequence of symbols with costs for things like vocabulary size and length of messages. All of these exciting avenues remain to be pursued in future work.

References

Barrett, Jeffrey A (2007). "Dynamic Partitioning and the Conventionality of Kinds". In: *Philosophy of Science* 74, pp. 527–546.

— (2009). "The Evolution of Coding in Signaling Games". In: *Theory and Decision* 67.2, pp. 223–237. DOI: 10.1007/s11238-007-9064-0.

Barrett, Jeffrey A, Brian Skyrms, and Calvin Cochran (2018). "Hierarchical Models for the Evolution of Compositional Language". In: *26th Philosophy of Science Association Biennnial Meeting.*

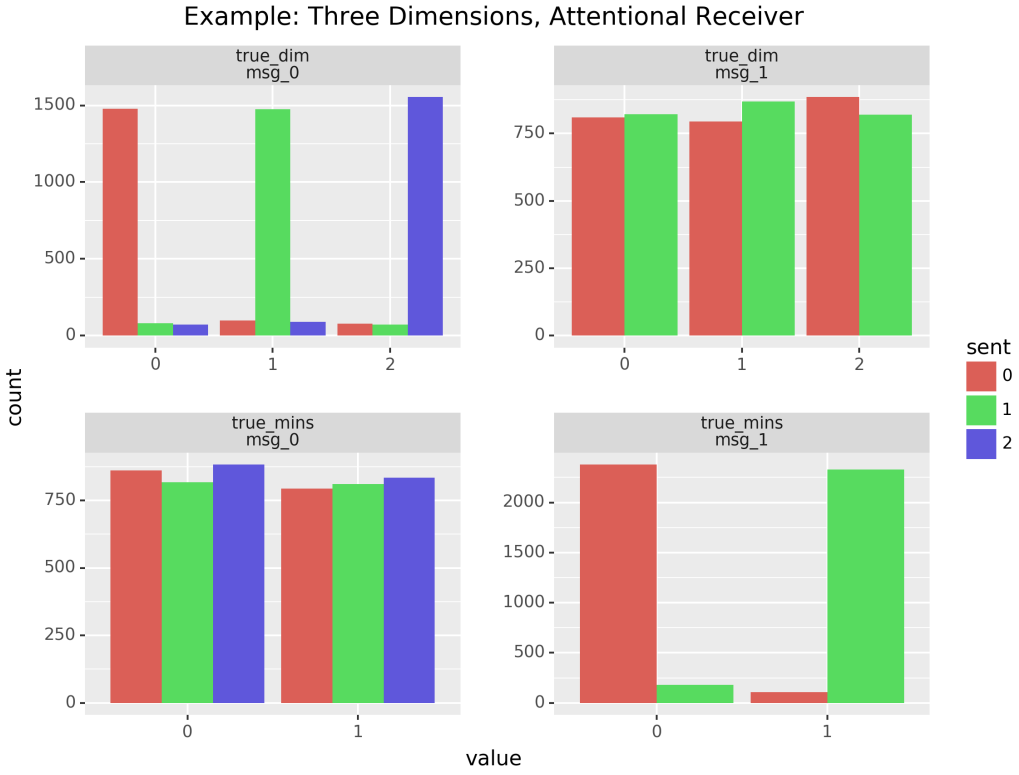


Figure 8: Example communication system with attentional receiver and three dimensions.

Berthet, Mélissa et al. (2018). "Titi monkey alarm sequences: when combining creates meaning". In: *26th Philosophy of Science Association Biennnial Meeting*.

Carnie, Andrew (2006). *Syntax: A Generative Introduction*. Second. Oxford: Blackwell Publishing.

Carroll, Lewis (1871). *Through the Looking-Glass, and What Alice Found There*. Macmillan.

Cäsar, Cristiane et al. (2013). "Titi monkey call sequences vary with predator location and type". In: *Biology Letters* 9.20130535, pp. 2–5. DOI: 10.1098/rsbl.2013.0535.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2016). "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". In: *International Conference of Learning Representations*. URL: http://arxiv.org/abs/1511.07289.

Franke, Michael (2016). "The Evolution of Compositionality in Signaling Games". In: *Journal of Logic, Language and Information*. DOI: 10.1007/s10849-015-9232-5.

Frege, Gottlob (1923). "Logische Untersuchungen. Dritter Teil: Gedankengefüge ("Compound Thoughts")". In: *Beiträge zur Philosophie des deutschen Idealismus* III, pp. 36–51. DOI: 10.1093/mind/LI.202.200.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. The MIT Press. URL: https://www.deeplearningbook.org/.

Graff, Delia (2000). "Shifting Sands: An Interest-Relative Theory of Vagueness". In: *Philosophical Topics* 28.1, pp. 45–81.

Heim, Irene and Angelika Kratzer (1998). *Semantics in Generative Grammar*. Blackwell Textbooks in Linguistics. Wiley-Blackwell.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

Jacobson, Pauline (2014). *Compositional Semantics: An Introduction to the Syntax/Semantics Interface*. Oxford Textbooks in Linguistics. Oxford Univeristy Press.

Janssen, Theo M V (1997). "Compositionality". In: *Handbook of Logic and Language*. Ed. by Johan van Benthem and Alice ter Meulen. Elsevier Science. Chap. 7, pp. 417–473. DOI: 10.1016/B978-044481714-3/50011-4.

Kaplan, David (1978). "Dthat". In: *Syntax and Semantics*. Ed. by Peter Cole. Vol. 9. New York: Academic Press, pp. 212–233.

Kennedy, Christopher (2007). "Vagueness and grammar: the semantics of relative and absolute gradable adjectives". In: *Linguistics and Philosophy* 30, pp. 1–45. DOI: 10.1007/s10988-006-9008-0.

Kennedy, Christopher and Louise McNally (2005). "Scale Structure, Degree Modification, and the Semantics of Gradable Predicates". In: *Language* 81.2, pp. 345–381.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference of Learning Representations (ICLR)*. arXiv: 1412.6980. URL: https://arxiv.org/abs/1412.6980.

Lazaridou, Angeliki, Alexander Peysakhovich, and Marco Baroni (2017). "Multi-Agent Cooperation and the Emergence of (Natural) Language". In: *International Conference of Learning Representations (ICLR2017)*. arXiv: 1612.07182. URL: http://arxiv.org/abs/1612.07182.

Levin, Beth and Malka Rappaport Hovav (2005). *Argument Realization*. Cambridge University Press.

Lewis, David (1969). *Convention*. Blackwell.

— (1988). "Relevant Implication". In: *Theoria* 54.3, pp. 161–174. DOI: 10.1111/j.1755-2567.1988.tb00716.x.

Mnih, Volodymyr et al. (2014). "Recurrent Models of Visual Attention". In: pp. 1–12. arXiv: 1406.6247. URL: http://arxiv.org/abs/1406.6247.

Mordatch, Igor and Pieter Abbeel (2018). "Emergence of Grounded Compositional Language in Multi-Agent Populations". In: *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*. URL: http://arxiv.org/abs/1703.04908.

Muysken, Pieter (2008). *Functional Categories*. Cambridge: Cambridge University Press.

Nielsen, Michael A (2015). *Neural Networks and Deep Learning*. Determination Press. URL: http://neuralnetworksanddeeplearning.com/.

Nowak, Martin A and David C Krakauer (1999). "The evolution of language". In: *Proceedings of the National Academy of Sciences* 96, pp. 8028–8033.

O'Connor, Cailin (2014). "Evolving Perceptual Categories". In: *Philosophy of Science* 81.5, pp. 840–851.

Ouattara, Karim, Alban Lemasson, and Klaus Zuberbühler (2009). "Campbell's monkeys concatenate vocalizations into context-specific call sequences." In: *Proceedings of the National Academy of Sciences* 106.51, pp. 22026–22031. DOI: 10.1073/pnas.0908118106.

Pagin, Peter and Dag Westerståhl (2010a). "Compositionality I: Definitions and Variants." In: *Philosophy Compass* 5.3, pp. 250–264. DOI: 10.1111/j.1747-9991.2009.00228.x.

— (2010b). "Compositionality II: Arguments and Problems." In: *Philosophy Compass* 5.3, pp. 265–282. DOI: 10.1111/j.1747-9991.2009.00229.x.

Partee, Barbara Hall (1995). "Lexical Semantics and Compositionality". In: *Invitation to Cognitive Science, Part 1: Language*. Ed. by Lila Gleitman and Mark Liberman. Cambridge: MIT Press. Chap. 11, pp. 311–360.

Rizzi, Luigi and Guglielmo Cinque (2016). "Functional Categories and Syntactic Theory". In: *Annual Review of Linguistics* 2.1, pp. 139–163. DOI: 10.1146/annurev-linguistics-011415-040827.

Roth, Alvin E. and Ido Erev (1995). "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term". In: *Games and Economic Behavior* 8, pp. 164–212.

Schlenker, Philippe, Emmanuel Chemla, Kate Arnold, et al. (2014). "Monkey semantics: two 'dialects' of Campbell's monkey alarm calls". In: *Linguistics and Philosophy* 37, pp. 439–501. DOI: 10.1007/s10988-014-9155-7.

Schlenker, Philippe, Emmanuel Chemla, Anne M Schel, et al. (2016a). "Formal monkey linguistics". In: *Theoretical Linguistics* 42.1-2, pp. 1–90. DOI: 10.1515/tl-2016-0001.

Schlenker, Philippe, Emmanuel Chemla, Anne M Schel, et al. (2016b). "Formal monkey linguistics: The debate". In: *Theoretical Linguistics* 42.1-2, pp. 173–201. DOI: 10.1515/tl-2016-0010.

Skyrms, Brian (2010). *Signals: Evolution, Learning, and Information*. Oxford University Press.

Steinert-Threlkeld, Shane (2014). "Learning to Use Function Words in Signaling Games". In: *Proceedings of Information Dynamics in Artificial Societies (IDAS-14)*. Ed. by Emiliano Lorini and Laurent Perrussel.

— (2016a). "Compositional Signaling in a Complex World". In: *Journal of Logic, Language and Information* 25.3, pp. 379–397. DOI: 10.1007/s10849-016-9236-9.

— (2016b). "Compositionality and competition in monkey alert calls". In: *Theoretical Linguistics* 42.1-2, pp. 159–171. DOI: 10.1515/tl-2016-0009.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: an introduction*. Second Edi. The MIT Press.

Syrett, K., C. Kennedy, and J. Lidz (2010). "Meaning and Context in Children's Understanding of Gradable Adjectives". In: *Journal of Semantics* 27.1, pp. 1–35. DOI: 10.1093/jos/ffp011.

Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine Learning* 8.3-4, pp. 229–256.

Xu, Kelvin et al. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *International Conference on Machine Learning (ICML 32)*. Ed. by Francis Bach and David Blei, pp. 2048–2057. arXiv: 1502.03044. URL: https://arxiv.org/abs/1502.03044.

Zuberbühler, Klaus (2018). "Combinatorial capacities in primates". In: *Current Opinion in Behavioral Sciences* 21, pp. 161–169. DOI: 10.1016/j.cobeha.2018.03.015.

A Full Experiment Details

For each number of dimensions n, a context has $2n$ objects. Each object is specified by n real numbers, chosen uniformly at random from the interval $(0, 2)$ at steps of 0.1. The values are uniformly subtracted by 1 to center them around 0.

The sender thus has $2n^2$ input nodes. As a convention, the first object for the sender is always the target. It has two hidden layers of 64 nodes each, with exponential linear activation.[29] The final hidden layer is then passed through two linear layers, with output sizes $|M_S|$ and 2, respectively. These are batch normalized[30] and fed into a softmax, to generate distributions over M_S and M_P.

The Basic Receiver receives the context, but with the objects in a random order compared to the sender, and two signals sampled from the sender's output distributions, encoded as one-hot vectors. It then has three rectified linear hidden layers of 64, 64, and 32 units respectively. Then a final linear layer with $2n$ output nodes (one for each target object) is passed through batch normalization and softmax to generate a distribution.

The Attentional Receiver passes the context and a messaged from M_S sampled from the sender through one exponential linear layer of 64 units, before batch normalization and softmax of size n, one for each dimension. A sample is taken from this distribution. The corresponding scalar values for each object along the dimension, together with a message sampled from the sender's distribution over M_P are passed through exponential linear layers of size 64 and 32, before batch normalization and softmax produce a distribution over target objects.

We trained using the REINFORCE algorithm, with mini-batches of size 64, and the Adam optimizer[31] with learning rate $5 \cdot 10^{-4}$. For $n = 1, 2, 3$ dimensions, and each type of receiver, we ran 10 trials of 5000, 20000, and 50000 mini-batches of training. After training, the trained networks then played 5000 versions of the game; the signals chosen, the target chosen, whether it was correct, and what the 'true' dimension and direction (min/max) for identifying the target in context were recorded.

Everything was implemented in PyTorch. The code and data are available at https://github.com/shanest/function-words-context.

[29] Clevert, Unterthiner, and Hochreiter (2016)
[30] Ioffe and Szegedy (2015)
[31] Kingma and Ba (2015)

A statistical learning approach to a problem of induction

Kino Zhao
yutingz3@uci.edu

University of California, Irvine
Logic and Philosophy of Science

(Draft updated December 7, 2018)

Abstract

At its strongest, Hume's problem of induction denies the existence of any well justified assumptionless inductive inference rule. At the weakest, it challenges our ability to articulate and apply good inductive inference rules. This paper examines an analysis that is closer to the latter camp. It reviews one answer to this problem drawn from the VC theorem in statistical learning theory and argues for its inadequacy. In particular, I show that it cannot be computed, in general, whether we are in a situation where the Vapnik-Chervonenkis (VC) theorem can be applied for the purpose we want it to.

Hume's problem of induction can be analyzed in a number of different ways. At the strongest, it denies the existence of any well justified assumptionless inductive inference rule. At the weakest, it challenges our ability to articulate and apply good inductive inference rules. This paper examines an analysis that is closer to the latter camp. It reviews one answer to this problem drawing from a theorem in statistical learning theory and argues for its inadequacy.

The particular problem of induction discussed in this paper concerns what Norton (2014) calls a formal theory of induction, where "valid inductive inferences are distinguished by their conformity to universal templates" (p.673). In particular, I focus on the template that is often called *enumerative induction*. An inductive argument of this type takes observations made from a small and finite sample of cases to be indicative of features in a large and potentially infinite population. The two hundred observed swans are white, so all swans are white. Hume argues that the only reason we think a

rule like this works is because we have observed it to work in the past, resulting in a circular justification.

Nevertheless, this kind of inductive reasoning is vital to the advancement of a scientific understanding of nature. Most, if not all, of our knowledge about the world is acquired through the examination of only a limited part of the world. The scientific enterprise relies on the assumption that at least some of such inductive processes generate knowledge. With this assumption in place, a weak problem of induction asks whether we can reliably and justifiably differentiate the processes that do generate knowledge from the ones that do not. This paper discusses this weak problem of induction in the context of statistical learning theory.

Statistical learning theory is a form of supervised machine learning that has not received as much philosophical attention as it deserves. In a pioneering treatment of it, Harman and Kulkarni (2012) argue that one of the central results in statistical learning theory – the result on Vapnik-Chervonenkis (VC) dimensions – can be seen as providing a new kind of answer to a problem of induction by providing a principled way of deciding if a certain procedure of enumerative induction is reliable. The current paper aims to investigate the plausibility of their view further by connecting results about VC dimension in statistical learning with results about NIP models in the branch of logic called model theory. In particular, I argue that even if Harman and Kulkarni succeed in answering the problem of induction with the VC theorem, the problem of induction only resurfaces at a deeper level.

The paper is organized as follows: section 1 explains the relevant part of statistical learning theory, the VC theorem, and the philosophical lessons it bears. Section 2 introduces the formal connection between this theorem and model theory and proves the central theorem of this paper. Section 3 concludes with philosophical reflections about the results.

1 Statistical learning theory

The kind of problems that is relevant for our discussion of VC dimensions is often referred to as classification problems that are irreducibly stochastic. In a classification problem, each individual is designated by its k-many features such that it occupies somewhere along a k-dimensional feature space, χ. The goal is to use this information to classify potentially infinitely many such individuals into finitely many classes. To

give an example, consider making diagnoses of people according to their test results from the k tests they have taken. The algorithm we are looking for needs to condense the k-dimensional information matrix into a single diagnosis: sick or not. The algorithm can be seen as a function $f : \chi \to \{0, 1\}$, where 1 means sick and 0 means not. For reasons of simplicity, I will follow the common practice and only consider cases of binary classification.

By "irreducibly stochastic", I mean that the target function f cannot be solved analytically. This might be because the underlying process is itself stochastic – it is possible for two people with exact same measures on all tests to nevertheless differ in health condition – or because the measurements we take have ineliminable random errors. This means that even the best possible f will make some error, and so the fact that a hypothesis makes errors in its predictions does not in itself count against that hypothesis. Instead, a more reasonable goal to strive towards is to have a known, preferably tight, bound on the error rate of our chosen hypothesis.

What makes this form of statistical learning "supervised learning" is the fact that the error bound of a hypothesis is estimated using data points whose true classes are known. Throughout this paper, I will use D to denote such a dataset. D can have any cardinality, but the interesting cases are all such that D is of finite size. Recall that the feature (or attribute) space χ denotes the space of all possible individuals that D could have sampled, so that $D \subset \chi$. I understand a hypothesis to be a function $h : \chi \to \{0, 1\}$. A set of hypotheses $\mathcal{H}$ is a set composed of individual hypotheses. Usually, the hypotheses are grouped together because they share some common features, such as all being linear functions with real numbers as parameters. This observation will become more relevant later.

One obvious way of choosing a good hypothesis from $\mathcal{H}$ is to choose the one that performs the best on D. I will follow Harman and Kulkarni (2012) and call this method enumerative induction, for it bears some key similarities with Hume's description of the observation of swans. This method is inductive because it has the ampliative feature of assuming that the chosen hypothesis will keep performing well on individuals outside of D. The question we are interested in is: how do we know this? What justifies the claim that the hypothesis performs well on D will perform well outside of D too? The answer that will be examined in this section and throughout the rest of the paper is that we know this claim to be true when we are in a situation where $\mathcal{H}$ has finite VC dimension, and the VC-theorem justifies this claim.

1. STATISTICAL LEARNING THEORY

To define the error rate of a hypothesis, recall the "ideal function" f mentioned in the introduction. Recall also that f classifies individuals from χ into $\{0, 1\}$, and f is imperfect. Nevertheless, since the process from χ to the classes is irreducibly stochastic, f is as good as we can hope for. Therefore, f will serve as our standard for the purpose of calculating the error rate of a hypothesis. Note that the hypotheses we are assessing are all from $\mathcal{H}$, our hypothesis set, but f need not be in $\mathcal{H}$.

Suppose D is of size N, and $x_1, \ldots, x_N \in D$. For each $h \in \mathcal{H}$ and $i \in [1, N]$, consider the random variable $X_i : \chi^N \to \{0, 1\}$ defined by

$$X_i\big(h(x_1, \ldots, x_N)\big) = \begin{cases} 1 & \text{if } h(x_i) \neq f(x_i), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Intuitively, $X_i = 1$ if the hypothesis we are evaluating, h, gives a different (and hence wrong) verdict on x_i than the target function f, and 0 otherwise. Assume $X_1, \ldots, X_N$ are independent, which is to say that making a mistake on one data point does not make it more or less likely for h to make a mistake on another one. This is typical if D is obtained through random sampling. Further assume $X_1, \ldots, X_N$ are identically distributed, which means that for any X_i and X_j in the sequence, $EX_i = EX_j$. This allows the error "rate" of h across multiple data points to be meaningfully computed.

Let $\overline{X} = \frac{1}{N}(\sum_{i=1}^{N} X_i)$, which is the measured mean error, and $\mu = E\overline{X}$, which is the expected mean error. I will follow Abu-Mostafa et al. (2012) in calling the former the *in-data error*, or E_{in}, and the latter *out-data error*, or E_{out}. To flesh out the relationship between these two values more clearly, we define

$$E_{in}(h) = \overline{X} = \frac{1}{N} \sum_{i=1}^{N} [\![h(\mathbf{x}_i) \neq f(\mathbf{x}_i)]\!] \tag{2}$$

$$E_{out}(h) = \mu = \mathbb{P}_N(h(\mathbf{x}) \neq f(\mathbf{x})) \tag{3}$$

Intuitively, the in-data error is the evidence we have about the performance of h, and the out-data error is the expectation that h will hold up to its performance. The amplification comes in when we claim that E_{out} is not very different from E_{in}. I will call the difference between E_{in} and E_{out} the *generalization error*.

For any single hypothesis, and for any error tolerance $\epsilon > 0$, Hoeffding (1963, p.16) proved a result called the *Hoeffding inequality* (see also Lin and Bai 2010, p. 70, and

Pons 2013, p. 205), which states that, under the assumption that the error rate for each data point is independent and identically distributed, we have (in the notations introduced above)

$$\mathbb{P}_N(|E_{in}(h) - E_{out}(h)| \geq \epsilon) \leq 2e^{-2\epsilon^2 N} \tag{4}$$

This inequation says that the probability of having a large generalization error in the assessment of a single hypothesis is bounded by $2e^{-2N\epsilon^2}$, which is a function of the size of the dataset, N, and the error tolerance ϵ.

Once we establish a bound in the case of a single hypothesis, we can get a similar bound for finitely many such hypotheses. The reason we cannot simply apply the Hoeffding inequality to our preferred hypothesis is that it requires us to pick a hypothesis before we compute its error rate from the data. This will not help us if we need to use data to do the picking. Instead, we need to make sure *any* hypothesis we pick out will have low enough generalization error, before we can trust the method (of enumerative induction) we use to pick.

Since we assume that the error rate of one hypothesis is independent of another, the probability of any of the finitely many hypotheses we are considering having a large generalization error is just going to be the union of the probability of each one of them does. In symbolic form, suppose there are $1 \leq M < \infty$ many hypotheses in $\mathcal{H}$, then we have

$$\mathbb{P}(\max_{h\in\mathcal{H}} |E_{in}(h) - E_{out}(h)| \geq \epsilon) = \mathbb{P}(\exists h \in \mathcal{H} | E_{in}(h) - E_{out}(h)| \geq \epsilon) \leq 2Me^{-2\epsilon^2 N} \tag{5}$$

While this bound may seem "loose", it serves our purpose when we have a reasonably small M or a reasonably large N.

This simple calculation becomes tricky, however, when $\mathcal{H}$ contains infinitely many hypotheses. If we replace M with infinity, then the upper bound stops being a bound, because $2Me^{-2\epsilon^2 N}$ grows to infinity as M does. This is where the VC dimension of $\mathcal{H}$ comes to play.

To understand the role of VC dimensions, define

$$\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \{(h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N)) \mid h \in \mathcal{H}\} \tag{6}$$

which is the set of all verdicts given by $\mathcal{H}$ on dataset D. If some hypotheses agree with each other on the classification of every data point, then their verdicts would be represented by the same tuple. This means that the cardinality of the set of verdicts

may be much smaller if $\mathcal{H}$ is very homogeneous. Moreover, different datasets of the same cardinality may elicit more or fewer different verdicts from $\mathcal{H}$. Define

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1,\cdots,\mathbf{x}_N \in \chi} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)| \tag{7}$$

as the max number of different verdicts $\mathcal{H}$ can generate from any dataset of cardinality N.

If all possible classifications of D have been represented in $\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)$, then we have $m_{\mathcal{H}}(N) = 2^N$. When this happens, we say that the hypothesis set $\mathcal{H}$ *shatters* the dataset D. Define the *VC dimension of* $\mathcal{H}$ to be the maximum N such that $m_{\mathcal{H}}(N) = 2^N$. In other words, it is the maximum number N such that there exists a dataset D of size N that is shattered by $\mathcal{H}$. If $m_{\mathcal{H}}(N) = 2^N$ holds for all N, then we say the VC dimension is infinite. Let's call a hypothesis set $\mathcal{H}$ VC-learnable if it has finite VC dimension.

Very roughly, the VC dimension of a hypothesis set tracks the maximum number of hypotheses that are still distinguishable from each other with respect to their verdicts on data. This means that, if we consider any more hypotheses, some of them will always agree with some others on all of the classifications they give to all possible data points, and so if one has low generalization error, the others will, too. The VC generalization bound is given as follows (Abu-Mostafa et al., 2012, p.53)

$$\mathbb{P}_N[\![\big(E_{out}(h) - E_{in}(h)\big) \leq \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}]\!] \geq 1 - \delta \tag{8}$$

where δ is the uncertainty tolerance. If $\mathcal{H}$ has an infinite VC dimension, then no such upper bound can be found. Notice that, holding everything else equal, increasing N brings the right-hand side down, which means that increasing data size allows us to make a better estimate of E_{out} with the same uncertainty tolerance. One can further show that

$$\lim_{N\to\infty} \mathbb{P}_N(\max_{h\in\mathcal{H}} |E_{in}(h) - E_{out}(h)| = 0) > 1 - \delta \tag{9}$$

for all $\delta > 0$. This means that, when $\mathcal{H}$ is either finite or has finite VC dimension, we can justifiably claim enumerative induction to be a reliable process that can pick out a good hypothesis from $\mathcal{H}$.

What makes this theorem especially powerful is not just that it shows how the error rates converge in the limit, but also that the convergence is uniform. What is

practically useful for statisticians is not so much that, if we have infinite data, we can figure out the true error rate of our hypothesis, but that, as soon as we know how many data points we have and the VC dimension of $\mathcal{H}$, we know precisely how confident we should be of our estimation of the error rate.

In what sense does this theorem answer a problem of induction? According to the analysis in Harman and Kulkarni (2012), this theorem defines precise conditions (i.e., ones where $\mathcal{H}$ has finite VC dimension) under which a particular inductive method (i.e., supervised learning in classification problems) is reliable. To the extent that we are concerned with the "easy" problem – the practical problem – of induction, the VC theorem does seem to provide a kind of answer we are looking for. In the next section, I challenge the applicability of this answer. In particular, I show that we can never know in general if we are in a situation where the above answer is applicable.

2 Finiteness of VC dimensions is uncomputable

A preliminary observation about the finiteness requirement is that we do not have a good grasp of what it means. What is the difference between these two sets of hypotheses such that one has finite VC dimension and the other does not? To put this point more concretely, we know that polynomial functions with arbitrarily high degrees have finite VC dimension, whereas the set of formulas with the sine function has infinite VC dimension. What is the difference between them? If we have a problem that can be reasonably formulated as polynomials or with a sine function, do we have good principled reasons why we should formulate it in one way rather than another?

Surprisingly, model theory in logic might help shed light on this question. It turns out that the concept of NIP theories corresponds to the class of hypothesis sets with finite VC dimensions. A theorem provably equivalent to the VC theorem was independently proved by the model theorist Shelah about these NIP theories and the corresponding NIP models. This connection was first recognized by Laskowski (1992). Interestingly, with the real numbers as their underlying domains, models with the usual plus and multiplication signs are NIP, whereas adding the sine curve makes them not NIP. This suggests that we can ask the same questions we would like to ask about our statistical hypothesis sets in model theory, which has a richer structure that is better understood independently.

In the previous section we discussed how the idea of "distinguishable hypotheses"

is important for the VC theorem. If a hypothesis set has finite VC dimension, we can think of it as having finitely many *distinguishable* hypotheses, even if it in fact has infinitely many. Intuitively speaking, if our dataset is "large enough" that not every combination of verdicts is representable with our hypotheses, then we can talk about which hypothesis is truly better than its competitors, as opposed to accidentally matching the specific data points. Having finite VC dimension ensures that there exist finite datasets that are "large enough". If a hypothesis set has finite VC dimension, let us call the set *VC-learnable*.

The corresponding concept in model theory relies on the same idea of distinguishability. Intuitively, if a formula is NIP – has the not-independent property – then there exists a natural number n such that no set larger than that number can be defined using this formula. A model is NIP just in case all of its formulas are (a formal definition is presented in Appendix A; for more formal details, see Simon, 2015).

We can then treat each hypothesis set as a formula defined on some domain. Laskowski (1992) shows that a hypothesis set is VC-learnable just in case the corresponding formula is NIP. What makes this correspondence especially useful is that model theorists have devoted a lot of efforts into determining which model is NIP. Once we know of a model that it's NIP, we also know that any hypothesis sets formulated using the language and domain of this model are VC-learnable.

For example, there is a group of models called *o-minimal*, which roughly means that all the definable subsets of the domain are finite unions of simple topological shapes like intervals and boxes. It suffices for our purposes to note that all o-minimal models are NIP (van den Dries, 1998, p. 90). As it happens, the real numbers with just addition and multiplication are o-minimal (van den Dries, 1998, p. 37). This means that any hypothesis set consisted of addition, multiplication, and the real numbers are going to have finite VC dimension. Similarly, the real numbers with addition, multiplication, and exponentiation is also o-minimal (Wilkie, 1996). This means that all sets of polynomials are VC-learnable.

As alluded to already, the real numbers with the sine function added are not NIP. This is roughly because, with the sine function, we can define copies of the integers using the set $\{x \in \mathbb{R} : \sin(x) = 0\}$, which allows us to define all of second-order arithmetic, and second-order arithmetic allows coding of arbitrary finite sets. As expected, this is reflected in statistical learning theory by the fact that the set of sine functions has infinite VC dimension, and so is not VC-learnable.

2. FINITENESS OF VC DIMENSIONS IS UNCOMPUTABLE

Another important observation from model theoretic investigations on NIP theory is that there seem to be no easy test for when an expansion of the real numbers is NIP. Although the relationship between the NIP property and properties like o-minimal and stable (a set of structures that are not o-minimal but are NIP) is well-researched and understood, there is no uniform way of telling where exactly a model lies (see, e.g., Miller, 2005[1]).

The statistical learning community echoes this difficulty with the observation that "it is not possible to obtain the analytic estimates of the VC dimension in most cases" (Shao et al., 2000; also see Vapnik et al., 1994). Recall that the VC dimension decides how big a dataset is "big enough". If the view is that enumerative induction is a reliable method when we are confident (i.e., low δ) that its assessment of hypotheses generalizes (i.e., low ϵ) and the VC theorem is supposed to guarantee this, then our inability to analytically solve the VC dimension of a given hypothesis set seems deeply handicapping.

To make the matter worse, it turns out that even knowing when we do have finite VC dimension is not a straightforward task, as witnessed by the following theorem, whose proof is given in Appendix A

Theorem 1. *The set $\{\varphi(x,y) : \varphi(x,y)$ is $NIP\}$, where $\varphi(x,y)$ is formulated in the language of arithmetic with addition and multiplication, is not decidable. In particular, this set computes $\emptyset^{(2)}$, the second Turing jump of the empty set.*

What this theorem tells us is that, in general, there is no effective procedure we can follow that can tell us, for any 2-place formula $\varphi(x,y)$, if it's NIP. With Laskowski's result, this means that we cannot compute, in general, if a given hypothesis set is VC-learnable either.

The specific way in which the set of all NIP formulas is uncomputable is significant also. For some time now, philosophers who study knowledge and learning from a formal perspective have placed a lot of emphasis on learning in the limit. Kelly (1996, p.52), for example, argues that the concept of knowledge (as opposed to, say, mere belief) implies that the method of generating such beliefs is stable in the limit. He then argues that the best way to formalize the notion of "stability in the limit" is to understand it as computable in the limit. Relatedly, a venerable tradition of formal learning

[1]Technically, Miller is interested in dichotomy theorems which establish either that an expansion of the reals is o-minimal or that it defines second-order arithmetic. As mentioned before, the former suffices for being NIP, and the latter suffices for being not NIP.

theory following Gold (1967) has explored extensively the conditions under which a noncomputable sequence may or may not be approximated by a computable sequence making only finitely many mistakes (cf. Osherson et al., 1986; Jain et al., 1999). From this perspective, it seems we might still be able to claim knowledge of what is or isn't knowable if we can compute the set of NIP formulas in the limit. Unfortunately, this latter task cannot be accomplished. This is because that, in order for a sequence to be approximable in the limit by another sequence, it cannot be harder than the first Turing jump of the sequence used to approximate it (Soare, 1987, p.57; see also Kelly, 1996, p.280). This means that something that is at least as hard as the second Turing jump cannot be approximated by a computable sequence.

To recapitulate the dialectic so far: an easy problem of induction asks us to identify and then justify the conditions under which a given ampliative method is reliable. The VC theorem gives one answer: supervised statistical learning from data is reliable just in case the hypothesis set has finite VC dimension. However, it turns out that we cannot, in general, decide if a hypothesis set is VC-learnable.

Can we judge our $\mathcal{H}$ on a case-by-case basis? Once we fix an $\mathcal{H}$, we can usually tell if it has finite VC dimension, and we can develop methods of empirically estimating its VC dimension using multiple datasets with varying sizes. However, this seems to just push the same problem to a deeper level. The problem that a method "sometimes is reliable, sometimes isn't", is solved by specifying a condition under which it always is reliable. But the problem that the condition "sometimes occurs, sometimes doesn't" seems to have no simple solution. In fact, the above theorem says that the latter problem has no solution.

3 Conclusion

A reasonable conclusion to draw from the discussions we've had so far, I think, is that the VC theorem still does not give us the kind of robust reliability we need to answer a question with some scope of philosophical generality. As is typical of answers people give to problems of induction, as soon as a rule is formulated, a question arises concerning its applicability. Similarly, what started out as a concern over the robustness of the method of enumerative induction turns into a concern over the robustness of the identifiable condition (i.e., the VC-learnable condition) under which enumerative induction is justified to be reliable.

A related question concerns the distinction, if there is one, between the cases where $\mathcal{H}$ has infinite VC dimension and cases where it has a VC dimension so large that it's impractical for us to make use of it. There is a sense in which the case of an infinite VC dimension fails *in principle*, whereas the case of a very large VC dimension only fails in *practice*. However, it is often impossible to analytically solve the VC dimension of a hypothesis set even if we do know that it's VC-learnable. Together with the result that we cannot test if a case is VC-learnable *in principle*, it seems to suggest that any information we might gain from the distinction between failing in principle and failing in practice will not be very informative, since we often can't tell which case we are in.

The philosophical difficulties discussed above raise an interesting question of how the practitioners view the same obstacle. Perhaps the way out is to accept a 'piecemeal' solution after all. It seems that when the VC dimension is small, we can often know both that it is finite, and that it is small. Theorists have also developed ways of estimating VC dimension using multiple datasets (see, e.g., Vapnik et al., 1994 and Shao et al., 2000). It seems that, as it often happens, philosophical problems are much more manageable when we do not look for principled solutions.

Acknowledgement

I would like to express my gratitude towards Sean Walsh for his supervision, as well as towards the participants in the 2016 Logic Seminar and attendees of the Society for Exact Philosophy 2017 meeting for their valuable feedback and discussion.

Appendix A

This appendix presents the proof of Theorem 1. I will follow the definition of NIP formulas given by Simon (2015) as follows (with notations changed to match preceding text)

> Let $\varphi(x;y)$ be a partitioned formula. We say that a set A of $|x|$-tuples is *shattered* by $\varphi(x;y)$ if we can find a family $(b_I : I \subseteq A)$ of $|y|$-tuples such that
>
> $$M \models \varphi(a;b_I) \iff a \in I, \quad \text{for all } a \in A$$

A formula $\varphi(x;y)$ is NIP if no infinite set of $|x|$-tuples is shattered by it.

3. CONCLUSION

Following notations from Soare (1987), let W_e to be the domain of the e-th partial recursive function and $Fin = \{e : W_e < \omega\}$.

Lemma Given e, define the following formula in the language of arithmetic

$$\begin{aligned}\theta_e(x,y) =& \exists l > x \; \exists \text{ enumeration } c_1, \ldots c_{2^l}, \text{ first } 2^l \text{ elements of } W_e \\ & \wedge \exists |\sigma| = l \text{ with } y = c_\sigma \wedge \sigma(x) = 1\end{aligned}$$

Then $e \in Fin$ iff θ_e is NIP.

Proof. ($\Rightarrow$) Suppose $e \in Fin$. The claim is: there is finite number N such that $|W_e| \leq 2^N$, and for all n, if a set A with cardinality n is shattered by θ_e, then $n \leq N$.

In particular, we show that the claim holds for N being the size of W_e. For the sake of contradiction, suppose there is A, with size n, shattered by θ_e, and $n > N$.

Let $A = \{a_1, \ldots, a_n\}$, $\{b_I : I \subset \{a_1, \ldots, a_n\}\}$, such that $\theta_e(a_i, b_I)$ iff $a_i \in I$.

Without loss of generality, let $a_n \geq n - 1$, and $I = \{a_n\}$. Then $a_n \in I$, and $\theta_e(a_n, b_I)$. This means that $\exists l > a_n \geq n - 1$ with the first 2^l many elements of W_e enumerated. Recall that the reductio hypothesis states $n > N$. This means that $|W_e| \geq 2^l > 2^{n-1} \geq 2^N$. This contradicts the original assumption that $|W_e| \leq 2^N$.

($\Leftarrow$) To show the contrapositive of this direction, suppose $e \notin Fin$, $|W_e| = \omega$. The claim is: θ_e is IP. Namely, $\forall N \; \exists n \geq N$, with some set A of cardinality n that is shattered by θ_e.

Take an arbitrary $n \geq N$. Let $A = \{0, \ldots, n-1\}$. Let b_σ's be the first 2^n elements of W_e, as σ ranges over finite strings of length n. Since σ is a string, we say $a \in \sigma \Leftrightarrow \sigma(a) = 1$.

We need to show that $\theta_e(a, b_\sigma) \Leftrightarrow \sigma(a) = 1$.

The left to right direction is trivial, since it is part of $\theta_e(a, b_\sigma)$ to state that $\sigma(a) = 1$.

To show the right to left direction, note that since $|W_e| = \omega$, there definitely exists an initial segment of 2^n many elements of W_e, and $n > a$ for all $a \in A$. This satisfies the first conjunct. To satisfy the second conjunct of θ_e, recall that we defined our enumeration to be such that $|\sigma| = n$ with σ being identified with every number $\leq 2^n$. This means that an enumeration of $c_1 \ldots c_{2^n}$ includes all c_σ with $|\sigma| = n$. Define $b_\sigma = c_\sigma$, and we are guaranteed that b_σ is in the enumeration, and $|\sigma| = n$. Finally, the last conjunct of θ_e is satisfied by supposition.

□

Theorem. *The set* $\{\varphi(x,y) : \varphi(x,y)$ *is* $NIP\}$*, where* $\varphi(x,y)$ *is formulated in the language of arithmetic with addition and multiplication, is not decidable. In particular, this set computes* $\emptyset^{(2)}$*, the second Turing jump of the empty set.*

Proof. Suppose not, then for any formula $\varphi(x,y)$, we can decide if it's NIP. This means that, for any e, we can decide if $\theta_e(x,y)$ as defined in the lemma above is NIP. By lemma, $\theta_e(x,y)$ is NIP just in case $e \in Fin$. If we could decide the former, we would be able to decide the set Fin. But by Soare (1987, p.66, Theorem 3.2), Fin is Σ_2-complete, and so computes $\emptyset^{(2)}$, the second Turing jump of the empty set, and hence is not computable. □

Bibliography

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook Singapore.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Harman, G. and Kulkarni, S. (2012). *Reliable reasoning: Induction and statistical learning theory*. MIT Press.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Jain, S., Osherson, D. N., Royer, J., and Sharma, A. (1999). *Systems that learn: an introduction to learning theory*. MIT press.

Kelly, K. T. (1996). *The logic of reliable inquiry*. Oxford University Press.

Laskowski, M. C. (1992). Vapnik-Chervonenkis classes of definable sets. *Journal of the London Mathematical Society*, 45(2):377–384.

Lin, Z. and Bai, Z. (2010). *Probability inequalities*. Science Press Beijing, Beijing; Springer, Heidelberg.

Miller, C. (2005). Tameness in Expansions of the Real Field. In *Logic Colloquium '01*, volume 20 of *Lecture Notes in Logic*, pages 281–316. Associaton for Symbolic Logic, Urbana, IL.

Norton, J. D. (2014). A material dissolution of the problem of induction. *Synthese*, 191(4):671–690.

Osherson, D. N., Stob, M., and Weinstein, S. (1986). *Systems that learn: An introduction to learning theory for cognitive and computer scientists.* The MIT Press.

Pons, O. (2013). *Inequalities in analysis and probability*. World Scientific.

Shao, X., Cherkassky, V., and Li, W. (2000). Measuring the VC-dimension using optimized experimental design. *Neural computation*, 12(8):1969–1986.

Simon, P. (2015). *A Guide to NIP Theories*. Lecture Notes in Logic. Cambridge University Press, Cambridge.

Soare, R. I. (1987). *Recursively Enumerable Sets and Degrees*. Perspectives in Mathematical Logic. Springer, Berlin.

van den Dries, L. (1998). *Tame Topology and O-Minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge.

Vapnik, V., Levin, E., and Le Cun, Y. (1994). Measuring the VC-dimension of a learning machine. *Neural computation*, 6(5):851–876.

Wilkie, A. J. (1996). Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094.