

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association

Atlanta, GA; 3-5 November 2016

Version: 29 October 2016

PhilSci
A · R · C · H · I · V · E



PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association
Atlanta, GA; 3-5 November 2016

This conference volume was automatically compiled from a collection of papers deposited in PhilSci-Archive in conjunction with PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association (Atlanta, GA; 3-5 November 2016).

PhilSci-Archive offers a service to those organizing conferences or preparing volumes to allow the deposit of papers as an easy way to circulate advance copies of papers. If you have a conference or volume you would like to make available through PhilSci-Archive, please send an email to the archive's academic advisors at philsci-archive@mail.pitt.edu.

PhilSci-Archive is a free online repository for preprints in the philosophy of science offered jointly by the Center for Philosophy of Science at the University of Pittsburgh, University Library System at the University of Pittsburgh, and Philosophy of Science Association

Compiled on 29 October 2016

This work is freely available online at:

<http://philsci-archive.pitt.edu/view/confandvol/confandvol2016PSA.html>

All of the papers contained in this volume are preprints. Cite a preprint in this document as:

Author Last, First (year). Title of article. Preprint volume for PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association, retrieved from PhilSci-Archive at <http://philsci-archive.pitt.edu/view/confandvol/confandvol2016PSA.html>, Version of 29 October 2016, pages XX - XX.

All documents available from PhilSci-Archive may be protected under U.S. and foreign copyright laws, and may not be reproduced without permission.

Table of Contents

	Page
Shahar Avin, <i>Centralised Funding and the Division of Cognitive Labour</i>	1
Massimiliano Badino, <i>How to Make Selective Realism More Selective (and More Realist Too)</i>	13
Sindhuja Bhakthavatsalam, <i>Duhemian Good Sense and Agent Reliabilism</i>	34
Brandon Boesch, <i>There Is A Special Problem of Scientific Representation</i>	50
Pierrick Bourrat and Qiaoying Lu, <i>Dissolving the missing heritability problem</i>	71
Thomas Boyer-Kassem, <i>Scientific expertise, risk assessment, and majority voting</i>	94
Carl Brusse and Justin Brunner, <i>Responsiveness and robustness in the David Lewis signalling game</i>	107
Ruey-Lin Chen and Jonathon Hricko, <i>Experimental Individuation and Retail Arguments</i>	118
M. Chirimuuta, <i>Crash Testing an Engineering Framework in Neuroscience</i> :	140
Alberto Cordero, <i>Eight Myths about Scientific Realism</i>	160
Wei Fang, <i>Concrete Models and Holistic Modelling</i>	174
Luke Fenton-Glynn, <i>Probabilistic Actual Causation</i>	194
Remco Heesen, <i>When Journal Editors Play Favorites</i>	212
Nicholaos Jones, <i>Strategies of Explanatory Abstraction in Molecular Systems Biology</i>	255

Michael Keas, <i>How the Diachronic Theoretical Virtues Make an Epistemic Difference.</i>	271
Adam Koberinski, <i>Reconciling axiomatic quantum field theory with cutoff-dependent particle physics.</i>	288
Soazig Le Bihan and Iheanyi Amadi, <i>Epistemically Detrimental Dissent: Contingent Enabling Factors v. Stable Difference Makers.</i>	308
Dennis Lehmkuhl, <i>Literal vs. careful interpretations of scientific theories: the vacuum approach to the problem of motion in general relativity.</i>	328
Johannes Lenhard, <i>Holism, or the Erosion of Modularity - a Methodological Challenge for Validation.</i>	348
Peter J. Lewis and Don Fallis, <i>Accuracy, conditionalization, and probabilism.</i>	370
C.D. McCoy, <i>Can Typicality Arguments Dissolve Cosmology's Flatness Problem?</i>	385
Thomas Moller-Nielsen, <i>Invariance, Interpretation, and Motivation.</i>	395
Elias Okon and Daniel Sudarsky, <i>Black Holes, Information Loss and the Measurement Problem.</i>	407
Jun Otsuka, <i>The Causal Homology Concept.</i>	421
Stéphanie Ruphy and Baptiste Bedessem, <i>Serendipity: an Argument for Scientific Freedom?</i>	443
S. Andrew Schroeder, <i>Using Democratic Values in Science: an Objection and (Partial) Response.</i>	460
Ayelet Shavit, Anat Kolombus, and Aaron M. Ellison, <i>Two Roads Diverge in a Wood: Indifference to the Difference Between 'Diversity' and 'Heterogeneity' Should Be Resisted on Epistemic and Moral Grounds.</i>	475
Bradford Skow, <i>Levels of Reasons and Causal Explanation.</i>	498

Quayshawn Spencer, <i>In Defense of the Actual Metaphysics of Race</i>	514
Veronica J Vieland, <i>Measurement of Statistical Evidence: Picking Up Where Hacking (et al.) Left Off</i>	531
Isaac Wiegman, <i>What Basic Emotions Really Are: Encapsulated or Integrated?</i>	551
John Zerilli, <i>Multiple realization and the commensurability of taxonomies</i>	576
Karen R. Zwier, <i>Interventionist Causation in Thermodynamics</i>	605

CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

ABSTRACT. Project selection by funding bodies directly influences the division of cognitive labour in scientific communities. I present a novel adaptation of an existing agent-based model of scientific research, in which a central funding body selects from proposed projects located on an epistemic landscape. I simulate four different selection strategies: selection based on a god’s-eye perspective of project significance, selection based on past success, selection based on past funding, and random selection. Results show the size of the landscape matters: on small landscapes historical information leads to slightly better results than random selection, but on large landscapes random selection greatly outperforms historically-informed selection.

Word count: 4359

INTRODUCTION

National funding bodies support much of contemporary science. The selection criteria for funding have gained increasing attention within philosophy of science (Gillies, 2008; O’Malley et al., 2009; Haufe, 2013; Lee, 2015). Meanwhile, there has been growing interest in model-based approaches to understanding the social epistemic activities of scientists (Kitcher, 1990; Strevens, 2003; Weisberg and Muldoon, 2009; Grim, 2009; Zollman, 2010). The current paper builds on previous modelling tools to explore the effects of centralised selection mechanisms on the division of cognitive labour and the ability of scientific communities to efficiently discover significant truths.

Science aims at discovering significant truths, i.e. not just any truths, but truths that will eventually contribute in a meaningful way to well-being (Kitcher, 2001). This is the justification for the public support of science, including basic science (Bush, 1945). Some funding terminology: scientific projects have high *impact* (ex post) if they result in significant truths; projects have high *merit* (ex ante) if they are predicted to have high impact.

Polanyi (1962) analysed merit as being composed of three components: scientific value, plausibility and originality. Polanyi notes an essential tension between plausibility and originality: the more original a project, the more difficult it is to evaluate its plausibility. Polanyi advocates selection by peer review as a conformist position, that sacrifices the occasional meritorious original project while ensuring all supported research projects are plausible, to “prevent the adulteration of science

2 CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

by cranks and dabblers” (p. 8). Gillies (2008, 2014) takes an opposing position, arguing that the cost of losing (infrequent) highly original and meritorious research is much greater than the cost of occasionally supporting implausible research that ends up being of low impact. As an alternative to peer review, Gillies advocates random selection. The tension between plausibility and originality is clearly relevant to questions of effective division of cognitive labour, and has direct links to science policy. This tension, and its complexity, is explored in this paper.

I will argue that the results of the simulations presented are both significant and surprising. The simulations show that, under reasonable parameter values for at least some fields of science, choosing projects at random performs significantly better, in terms of accumulated significant truths, compared to other funding strategies, including project selection by peer review. The results support, to an extent, Gillies’ proposal of funding by lottery.

1. MODEL DESCRIPTION

The model explores the influence of different funding mechanisms on the accumulation of significant truths. It builds on the epistemic landscape model developed by Weisberg and Muldoon (2009), extending it by adding representations of centralised funding selection and dynamic changes in project merit. The latter is added to reflect a more realistic picture of scientific merit. For example, Strevens (2003) discusses the effect of a successful discovery on all further pursuits of the same question: they no longer have any merit, as they lose all originality. Several dynamic processes affecting merit are detailed later in the paper.

The model represents a population of scientists exploring a topic of scientific interest. They are all funded by the same central funding body to pursue projects of varying duration, measured in years. Each project’s significance is allocated in advance by the modeller, from a “god’s-eye” perspective. When grants end scientists successfully complete their project. Their projects’ results contribute to the collection of significant truths in the field’s corpus of knowledge. Funding mechanisms are compared by their ability to generate this accumulation of significant truths.

For simplicity, scientists in the model (unrealistically) do not share their findings nor explore similar projects during research. They only work on the project for which they were funded and they only share their results at the end of a grant. The social processes set aside here have been explored in previous works (Grim, 2009; Zollman, 2010). Future work may combine the different models towards a unified picture of the division of cognitive labour.

CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR 3

Funding is represented as a process of selection. In every time step, the scientists whose grants have run out are placed in a pool of candidates along with new entrants to the field, and the modelled funding mechanism selects from this pool of candidates those who will receive funding and carry out research projects. Modelled funding mechanisms differ in the way they select individuals, as outlined below.

Actual potential: Actual potential, which can only be known from a god's-eye perspective, is the significance of a project's results *were it successfully completed today*. In the absence of time-dependant merit, actual potential is simply the significance of the project's results. However, in the presence of time-dependence the significance could change between the initiation of the project (at the point of funding) and its completion (at the point of contributing the results to the relevant corpus). This means that in the presence of time-dependence, actual potential might diverge from the eventual contribution of the project.

Estimated potential: Estimated potential is the scientific community's ex ante evaluation (assumed, for simplicity, to be single-valued) of the merit of a proposed project. This prediction is taken to rely on the known contributions of past projects which bear some similarity to the proposed project, and so depends on the history of research projects in the field. In representing decisions based on the research community's prediction, this selection method is akin to peer-review.

Past funding: Under this mechanism, funding is allocated to those scientists who already received funding in the past, and only to them. The model (unrealistically) represents all scientists as being of equal skill, and so this mechanism cannot be taken to mean the selection of the most "intrinsically able" scientists. Rather, this mechanism is included as a "most conservative" option, not admitting any new researchers to the field beyond the field's original investigators.

Lottery: Under a lottery, all candidates have equal chances of being funded. The lottery option serves both as a natural benchmark for other funding methods, and as a representation of the mechanism proposed by Gillies (2014).

The essence of the model is the comparison of the performance of these selection mechanisms in generating results of high significance over time under various conditions.

To represent in the model the time-dependence of merit, the significance contributions of different project results are allowed to change over time as a response to scientists' actions. Three dynamic processes are included in the model (details in §2.5). Two processes involve a reduction of significance following a successful project or breakthrough,

4 CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

which reflects the one-off nature of discovery. The third process involves an increase in significance when a new avenue of research is opened by a significant discovery. Simulations based on the model show that these dynamic processes have a significant effect on the relative performance of different funding strategies.

2. SIMULATION DETAILS

2.1. Simulating the epistemic landscape. To investigate the complex nature of the domain being modelled, the model was turned into a computer simulation.¹ The basic structure of the landscape simulation follows Weisberg and Muldoon's, of a two-dimensional configuration space, charted with two coordinates x and y , with an associated scalar field represented in a third dimension as height along the z axis. Each (x, y) coordinate pair specifies a different potential research project; the closer two projects are on the landscape, the more similar they are. The scalar value associated to the coordinate represents the significance of the result obtained on a successful completion of the project, were it completed today (allowing for time dependence). The limit to two spatial dimensions of variation between projects is likely to be unrealistic (Wilkins, 2008), but a higher-dimensional alternative would make the model much less tractable.

In each run of the simulation, the landscape is generated anew in the following process:

- (1) Initialise a flat surface of the required dimensions.
- (2) Choose a random location on the surface.
- (3) Pick random values for relative height, width along x , and width along y .
- (4) Add to the landscape a hill at the location chosen in step 2 by using a bivariate Gaussian distribution with the parameters picked in step 3.
- (5) Repeat steps 2-4 until the specified number of hills is reached.
- (6) Scale up linearly the height of the landscape according to the specified maximum height.

This process generates the “god’s-eye” perspective of the research potential of the domain. Here and later, random variables are used to fill-in parameters whose existence is essential for the simulation, but where (1) the specific values they take can vary across a range of valid model targets, and/or (2) there is no compelling empirical evidence to choose a particular value. This requires, however, several runs of the simulation for each configuration, to average out the effects of random variation.

¹Source code for the simulation is available from the author on request.

2.2. Simulating agents. The agents in the model represent scientists investigating the epistemic landscape. Each agent represents an independent researcher or group, and is characterised by its location on the landscape, representing the project they are currently pursuing, and a countdown counter, representing the time remaining until their current project is finished. Like Weisberg and Muldoon’s “hill climbers”, agents are simulated as local maximisers. Agents follow the following strategy every simulation step:

- (1) Reduce countdown by 1.
- (2) If countdown is not zero: remain in same location.
- (3) If countdown is zero: contribute to the accumulated significance the significance of the current location, and attempt to move to the highest local neighbour.

In the simulation, the agents are identical, in the sense that any agent, when successfully completing a project of a given significance, will contribute exactly that amount to the accumulated significance of the field. This simplification ignores natural ability and gained experience, and stems from a focus on a particular approach to science funding, which funds *projects*, rather than funding *people*. The focus is informed by the explicit policies of certain funding bodies, like the National Institutes of Health (NIH), reflected, for example, in the institution of blind peer review. Thus, the results of the current work would not extend to the minority of science funding bodies, such as the Wellcome Trust, that make explicit their preference to fund people rather than projects.

The *local neighbourhood* of an agent is defined as the 3×3 square centred on their current position. The attempt to move to the highest neighbour depends on the selection (funding) mechanism, as discussed below. The *accumulated significance*, which is the sum of all individual contributions to significance, is stored as a global variable of the simulation and used to compare strategies.

In the beginning of the simulation, a specified number of agents are seeded in random locations on the landscape, with randomly generated countdowns selected from a specified range of values. An example of an initial seeding of agents can be seen in Fig. 1.

In the absence of selection and time-dependence, the course of the simulation is easy to describe: agents begin in random locations on a random landscape, and as the simulation progresses the agents finish projects and climb local hills, until, after an amount of time which depends on the size of the landscape, the number and size of peaks, and the duration of grants, all agents trace a path to their local maxima and stay there. Since agents increase their local significance during the climb, the rate of significance accumulation increases initially, until all agents reach their local maxima, at which point significance continues accumulating at a fixed rate indefinitely. This is the dynamic

6 CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

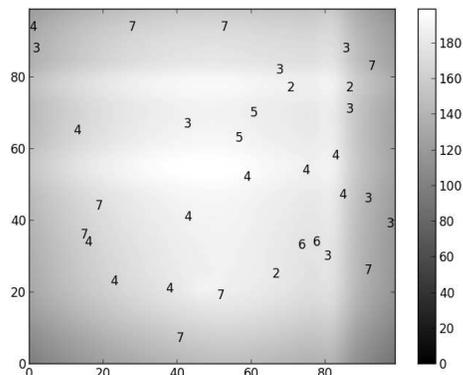


FIGURE 1. Landscape simulation with initial seeding of agents. Each number on the landscape represents an agent at its location, with the value of the number representing the agent’s countdown. The colours indicate the height (significance) of each position (project) in the landscape.

seen in Weisberg and Muldoon’s simulation for a pure community of “hill climbers”, and its unrealistic nature highlights the importance of simulating the time-dependence of significance.

2.3. Simulating communal knowledge. In addition to their contribution to significance, agents also contribute to the *visibility* of the landscape (Muldoon and Weisberg, 2011). The visibility of a project represents whether the scientific community, and especially funding bodies, can estimate the significance contribution of that project. Initially, the entire landscape is invisible, representing full uncertainty. Upon initial seeding of agents, each agent contributes vision of their local neighbourhood, as defined above, to the total vision. As the agents move, they add vision of their new local neighbourhood. Visibility is used in the *best_visible* funding mechanism described below.

The simulation represents visibility in a simplistic manner by assigning binary values: either the community knows what the significance of a project will be, or it does not. A more realistic representation will allow partial visibility, with some distance decay effect, such that the community would still be able to make predictions of significance for less familiar projects, but these predictions will have a probability of being wrong, with the probability of error increasing the more unfamiliar these projects are. This addition, however, will be computationally heavy, as it requires maintaining multiple versions of the landscape, both for the real values and for the estimated values.

2.4. Simulating funding strategies. The aim of the model is to explore the effects of funding mechanisms on the population and distribution of investigators. Since the aim is to simulate current funding practices (albeit in a highly idealised manner), and since current funding practices operate in passive mode (choosing from proposals originating from scientists rather than dictating which projects ought be pursued), the guiding principle of the simulation is that a funding mechanism is akin to a selection process: at each step of the simulation, the actual population of agents is a subset of the candidate or potential population, where inclusion in the actual population follows a certain selection mechanism.

Funding mechanisms are simulated in the following manner:

Every step:

- (1) Place all agents with zero countdown in a pool of “old candidates”.
- (2) Generate a set of new candidate agents, in a process identical to the seeding of agents in the beginning of the simulation.
- (3) Select from the joint pool of (old candidates + new candidates) a subset according to the selection mechanism specified by the funding method.
- (4) Only selected agents are placed on the landscape and take part in the remainder of the simulation, the rest are ignored.

The simulation can represent four different funding mechanisms:

best: selects the candidates which are located at the highest points, regardless of the visibility of their locations. This simulates a mechanism which selects the most promising projects from a god’s eye perspective. This overly optimistic mechanism does not represent a real funding strategy. Rather, it serves as an ideal benchmark against which realistic funding mechanisms are measured.

best_visible: filters out candidates which are located at invisible locations, i.e. candidates who propose to work on projects which are too different from present or past projects. It then selects the candidates in the highest locations from the remainder. This strategy is closer to a realistic representation of selection by peer review. Note that even this version is epistemically optimistic, as it assumes the selection panel has successfully gathered all available information from all the different agents, both past and present.

lotto: selects candidates at random from the candidate pool, disregarding the visibility and height of their locations.

oldboys: represents no selection: old candidates continue, no new candidates are generated.

8 CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

The key parameters for all funding mechanisms are the size of the candidate pool and the size of the selection pool. The size of the candidate pool, which in turn depends on the size of the new candidate pool (as the size of the old candidate pool emerges from the simulation), has been chosen in the simulations such that the total candidate pool is equal in size to the initial number of agents (except *oldboys* where there are no new candidates). This means the success probability changes between funding rounds, around a mean which is equal to $1/(\text{average countdown})$. With an average grant duration of five years, this yields a success rate of 20%, close to the real value in many contemporary funding schemes (NIH, 2014). The number of grants awarded each year is set to equal the number of grants completed each year, maintaining a fixed size for the population of investigators.

For simplicity, the simulated funding mechanisms do not take into account the positions of existing agents on the landscape, except indirectly when considering their vision. Future simulations may consider a selection mechanism which explicitly favours either diversity or agglomeration, though one expects difficulties in operationalisation and measurement of epistemic diversity.

2.5. Simulating merit dynamics. To make the simulation more realistic, the significance of projects is allowed to change over time in response to research activities of the community of investigators. Three such dynamic processes are included in the simulation:

Winner takes it all: As was made explicit by Strevens (2003), the utility gain of discovery is a one-off event: the first (recognised) discovery of X may greatly contribute to the collective utility, but there is little or no contribution from further discoveries of X. In the simulation, this is represented by setting the significance of a location to zero whenever an agent at that location has finished their project and made their contribution to accumulated significance. This effect is triggered whenever any countdown reaches zero, which makes it quite common, but it has a very localised effect, only affecting the significance of a single project.

Reduced novelty: When a researcher makes a significant discovery, simulated by finishing a project with associated significance above a certain threshold, the novelty of nearby projects is reduced, which in the model is simulated by a reduction of significance in a local area around the discovery.

New avenues: When a researcher makes a significant discovery, it opens up the possibility of new avenues of research, simulated in the model by the appearance of a new randomly-shaped hill at a random location on the landscape.

3. RESULTS AND DISCUSSION

Here I present the results of simulations of different setups of interest, exploring the relative success of different funding mechanisms under different conditions.

All simulation results show a comparison between the four funding mechanisms, as a plot of total accumulated significance (arbitrary units) at the end of the simulation run, averaged over five runs with different random seeds. In all simulations the range of countdowns was 2 to 7. The number of individuals was set to equal (size of landscape)^{3/4}. Simulations were ran for 50 steps. The trigger for significance-dependant processes was 0.7 of the global maximum. Results are shown for a small landscape (50×50) in Fig. 2 and for a large landscape (500×500) in Fig. 3.

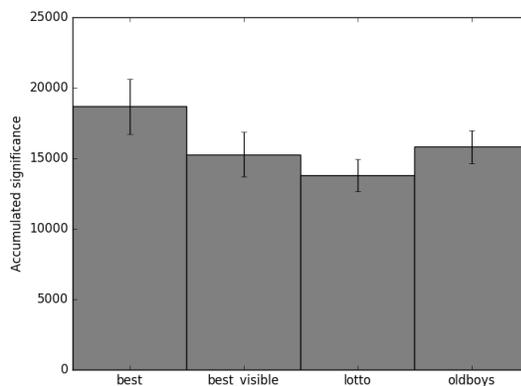


FIGURE 2. Comparison of significance accumulation under different funding mechanisms, small landscape (50×50).

To get a feeling for how the community is affected by the funding mechanism, I present visualisations of the state of the landscape at the end of the simulation run for the two funding mechanisms mentioned in the introduction (*best_visible* and *lotto*) in Fig. 4. Note that due to the *winner takes it all* dynamic process it is possible to “see” the past trajectory of exploration, as completed projects leave behind highly localised points of zero (remaining) significance. This allows for a visual representation of the division of cognitive labour that emerges under different funding schemes.

As is clear from the simulations, the *best* funding mechanism is indeed best at accumulating significance over time, though with various lead margins over the second best strategy. In the presence of dynamic

10 CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

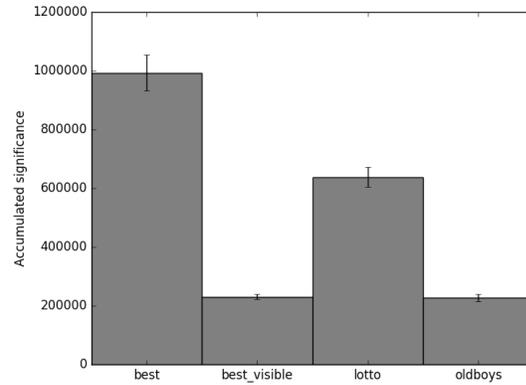


FIGURE 3. Comparison of significance accumulation under different funding mechanisms, large landscape (500×500).

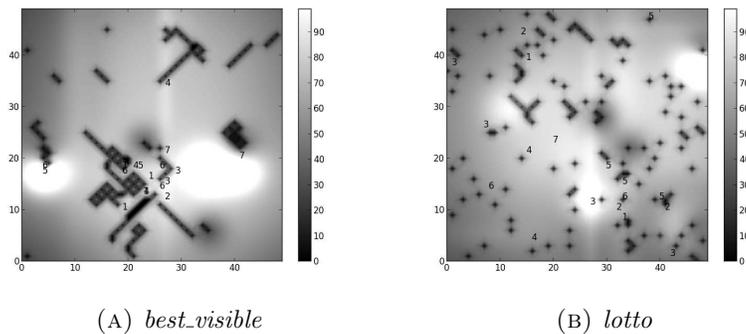


FIGURE 4. Landscape visualisation at the end of the simulation run under different funding mechanisms.

processes, *best* is in the best position to locate new avenues for research, wherever they show up. However, as mentioned above, the *best* funding strategy is not realisable, as it requires a god's eye view of the epistemic landscape.

On the small landscape the three strategies, *best_visible*, *oldboys*, and *lotto* perform roughly similarly, with *lotto* at a small disadvantage as it cannot make use of valuable information from past successes. It seems counter-intuitive that *best_visible* performs worse than *oldboys*. A possible explanation is the effect of reduced novelty: *best_visible* tends to cluster scientists around the most promising projects, and so when one makes a breakthrough it reduces the significance of contributions for all groups working on similar projects (the phenomenon known in

CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR 11

contemporary science as “scooping”). This excessive clustering around fashions is not present in *oldboys* or *lotto*.

On the large landscape *lotto* greatly outperforms *best_visible* and *oldboys*. This is because new avenues on a large landscape are likely to spawn outside the visibility of the agents, where *lotto* can access them but the other two strategies cannot. In the smaller landscape this effect is not apparent, as the relative visibility is larger, and therefore the chance of a new avenue appearing within the visible area is larger.

CONCLUSION

This paper presented a way to extend existing epistemic landscape models so that they can represent selection by a central funding body and time dependence of significance. This model was used in computer simulations to compare the effectiveness of different idealised versions of selection criteria, most notably selection based on past successes (akin to peer review), random selection and no selection. The most significant result from the simulation was that on a large landscape, when a topic can be explored in many ways that could be very different from each other, random selection performs much better than selection based on past performance.

This result fits in with a general result from the body of works on agent-based models of scientific communities, that shows diversity in the community trumps individual pursuit of excellence as a way of making communal epistemic progress. The tension of science funding, between originality and plausibility, is thus a part of the broader tension between diversity and excellence, between exploration and exploitation.

Previous social epistemology models have focused on the role of *internal* factors in shifting the balance between exploration and exploitation. Kitcher (1990); Strevens (2003) look at reward structures (of internal credit, not external monetary rewards) and individual motivation towards credit or truth. Grim (2009); Zollman (2010) look at information availability and information transfer between scientists, and at individual beliefs. Weisberg and Muldoon (2009) look at individual researchers’ social strategy: follower or maverick.

The current work is the first within this modelling lineage to look at the effects of an *external, institutional* factor: selection by a centralised funding body. The current paper brings this line of research closer to having a direct relevance to science policy. Hopefully future work in this vein will continue this trend, to deliver on the challenge set out by Kitcher (1990, p. 22):

How do we best design social institutions for the advancement of learning? The philosophers have ignored the social structure of science. The point, however, is to change it.

12 CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

We could start by advocating for funding mechanisms that allow for more exploration.

REFERENCES

- Bush, V. (1945). *Science, the endless frontier: A report to the President*. Washington: U.S. Government printing office.
- Gillies, D. (2008). *How should research be organised?* London: College Publications.
- Gillies, D. (2014). Selecting applications for funding: why random choice is better than peer review. *RT. A Journal on research policy and evaluation* 2(1).
- Grim, P. (2009). Threshold phenomena in epistemic networks. In *Complex adaptive systems and the threshold effect: Views from the natural and social sciences: Papers from the AAAI Fall Symposium*, pp. 53–60.
- Haufe, C. (2013). Why do funding agencies favor hypothesis testing? *Studies in History and Philosophy of Science Part A* 44(3), 363–374.
- Kitcher, P. (1990). The division of cognitive labor. *The journal of philosophy* 87(1), pp. 5–22.
- Kitcher, P. (2001). *Science, truth, and democracy*. New York: Oxford University Press.
- Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science* 82(5), 1272–1283.
- Muldoon, R. and M. Weisberg (2011). Robustness and idealization in models of cognitive labor. *Synthese* 183(2), 161–174.
- NIH (2014). Success rates - NIH research portfolio online reporting tools (RePORT). http://report.nih.gov/success_rates/, Accessed 11 July 2014.
- O'Malley, M. A., K. C. Elliott, C. Haufe, and R. M. Burian (2009). Philosophies of funding. *Cell* 138(4), 611–615.
- Polanyi, M. (1962). The republic of science: Its political and economic theory. *Minerva* 1, 54–73.
- Strevens, M. (2003). The role of the priority rule in science. *The journal of philosophy* 100(2), 55–79.
- Weisberg, M. and R. Muldoon (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of science* 76(2), 225–252.
- Wilkins, J. S. (2008). The adaptive landscape of science. *Biology and philosophy* 23(5), 659–671.
- Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis* 72(1), 17–35.

How To Make Selective Realism More Selective (and More Realist Too)

Massimiliano Badino

Massachusetts Institute of Technology — Universitat Autònoma de Barcelona

Abstract

Selective realism is the thesis that some wisely chosen theoretical posits are essential to science and can therefore be considered as true or approximately true. How to choose them wisely, however, is a matter of fierce contention. Generally speaking, we should favor posits that are effectively deployed in successful prediction. In this paper I propose a refinement of the notion of deployment and I argue that selective realism can be extended to include the analysis of how theoretical posits are actually deployed in symbolic practices.

1. Introduction

Among the several forms of realism, the so-called selective realism (SelRealism) is arguably the one that engages history of science more seriously. The driving idea of SelRealism is that, although theories as wholes are false and doomed to be abandoned, it is possible to select a certain number of theoretical posits (TPs) that are likely to be maintained in future theories and are therefore true or approximately true. How to determine these TPs is *partly* an empirical question—and this explains the historical character of the SelRealism program—but it cannot be *merely* an empirical question lest one end up in post-hoc rationalizations. A central issue of

SelRealism, hence, is how to specify criteria to properly conceptualize the TPs on which one should place one's realist commitment.

In this paper, I argue that contemporary approaches to SelRealism have neglected an important element related to the way in which theoretical claims are deployed in scientific theories (Section 2). In Section 3, I propose a refinement of SelRealism based on the distinction between deploying a TP fundamentally and deploying it in a non-accidental fashion. I use the concept of symbolic practices to articulate this distinction. Finally, in Section 4, I clarify my points by discussing the early development of perturbation theory.

2. Selective Realism: Theory and Practice

The upholders of SelRealism cherish two fundamental ambitions. First and foremost, they aim at making a good use of the so-called no-miracles argument (NMA) according to which one can justifiably infer the truth (or the approximate truth) of a successful theory, because, otherwise, the success would remain inexplicable. The NMA is considered to be the strongest support to realisms of any sort (Musgrave 1988; Psillos 1999, 68-94). A challenging objection to the NMA is the pessimistic meta-induction (PMI) originally formulated by Larry Laudan. According to this argument, the success of a theory is never a sufficient reason to infer even its approximate truth because history of science is replete with examples of very successful theories that wound up overthrown at some later stage. As it is likely the case that our most successful theories will suffer the same fate in the future, one has

to conclude that the realist commitment is not justified (Laudan 1981). Among the several responses to the PMI, one consists in noticing that the failures of past theories, in fact, did not depend on those TPs that lead them to success. In other words, granted Laudan's point that successful past theories are false as wholes, it can still be argued that the constituents of those theories that were responsible for their empirical success have been retained in our current science. Thus, the realist needs only to shift her commitment from theories as wholes to those enduring TPs that, being essential for success, can be justifiably believed to be true or approximately true.

The next question is, of course, how to determine those TPs. Thus, the second ambition of the upholders of SelRealism is to solve the problem of selectivity in some principled way and so beat the PMI. In one of the first instantiations of SelRealism, Philip Kitcher argued that one must "distinguish between those parts of theory that are genuinely used in the success and those that are idle wheels" (Kitcher 1993, 143). The point of this distinction is that credit for the success of a theory should be due only to those TPs that effectively contribute to it. Elaborating on Kitcher's intuition, one can argue that the program of SelRealism is based on two major conditions:

(S) Success condition: the selection of the important TPs must hinge on their relation with some significant success of the theory.

(D) Deployment condition: one must select those TPs that were effectively used in scoring that success.

Let me briefly comment on these two conditions. While (S) is now a realist trademark, the deployment condition (D) is what sets apart SelRealism from other forms of realism, such as structural realism, also engaged in picking out enduring elements of scientific theories (Worrall 1989; Chakravartty 2011). It is also important to notice that (S) and (D) are independent conditions. Firstly, (S) refers to a relation between the selected TP and empirical success, while (D) refers to a relation between the TP and the rest of the theory. Secondly, either condition can be satisfied separately. (D) has been added precisely to avoid those cases in which idle TPs are involved in empirical success and, obviously, there are scores of examples of TPs used by theories which however never led to any success. It follows that, while (S) is supposed to meet the first ambition of SelRealism, the second ambition, to block the PMI, is on (D).

So much for SelRealism in theory. Let us now examine how this program has been carried out in practice. One of the first philosophers to seriously elaborate on Kitcher's suggestion was Stathis Psillos. His criterion for selecting TPs works in the following way (Psillos 1999, 110). Let us assume that a certain successful prediction P can be obtained by combining the TPs H, H' and the auxiliaries A .¹ According to

¹ For virtually all writers, empirical success means "successful prediction". David Harker has leveled important criticisms against this tendency to interpret success in terms of individual predictions and has suggested that success should be understood as progress, i.e. in terms of the improvements a theory makes with respect to its predecessors (Harker 2008, 2013).

Psillos, the TP H is essential to success P and should be considered true or approximately true if and only if:

(1) H' and A alone do not lead to P .

(2) There is no alternative H^* to H such that:

(a) H^* is consistent with H' and A ;

(b) H^* , H' , and A lead to P ;

(c) H^* is not *ad hoc* or otherwise purposefully concocted to lead to P .

This criterion is the bedrock of Psillos's *divide et impera* strategy. The driving intuition behind it is to capture the *indispensability* of H : we should place our realist commitment upon those TPs without which empirical success cannot be obtained. However, Tim Lyons has cogently argued that Psillos's criterion fails to characterize indispensability (Lyons 2006). The indispensability of H should be ensured by condition (2), which states, in brief, that H cannot be replaced by any other TP. But, Lyons notices, "there will always be other hypotheses, albeit some that we find very unappealing, from which any given prediction can be derived" (Lyons 2006, 540). More importantly, Lyons argues, Psillos's criterion is not even an effective means for credit attribution, because it does not tell us much about how H contributes to the empirical success P . In particular, condition (2) has no relevance whatsoever for H 's specific contribution, because it only concerns conceivable alternatives to H , alternatives that, if H is at hand, nobody would even bother to explore. Lyons

perceptively stresses that the problem with Psillos's criterion boils down to the fact that it obliterates condition (D): "by introducing his criterion, [Psillos] has discarded the central idea of deployment realism—introduced by Kitcher and seemingly advocated by Psillos himself" (Lyons 2006, 541). It is interesting to note that, by dropping condition (D), Psillos's position becomes vulnerable to another form of PMI. One could think of getting around of Lyons's first objection by arguing that, even though an alternative to *H* is always conceivable, *at the present state* of our knowledge it is not, therefore the objection is empty. In other words, one could inject the time factor in Psillos's criterion and make it a statement of our actual best knowledge. But then the PMI crops up again, because history shows that there is no guarantee that what is indispensable today will be so tomorrow. The whole point of the PMI is that there is nothing special in our knowledge as far as it is considered *present*, because there have been a lot of *present knowledges* that have been blissfully abandoned. This is why one needs condition (D): what makes our present knowledge so special is not its happening at a certain time, but its having gone through a certain *process*, i.e., a form of deployment. The fact that our present knowledge has been deployed at lengths and it is still with us constitutes a reason to believe that it is true or approximately true.

3. Deconstructing Deployment

Having grasped that the flaw in Psillos's criterion is the dropping of the deployment condition, Lyons suggests to run to the other end of the spectrum and to inflate

dramatically the notion of deployment. His “responsibility model” consists in discarding selectivity altogether and in considering responsible for the empirical success of a theory each and every element that was originally deployed: “credit will have to be attributed to all responsible constituents, including mere heuristics (such as mystical beliefs), weak analogies, mistaken calculations, logically invalid reasoning etc.” (Lyons 2006, 543). Clearly, Lyons’s proposal amounts to a crack-up of the entire SelRealism program. But, more importantly, I do not think that the responsibility model captures the correct significance of (D). As my previous considerations about the PMI show, the deployment condition is not merely supposed to tell us that a TP has been effectively used in obtaining empirical success (as opposed to be *dispensable*), but also that it has been robustly so (as opposed to be merely *accidental*). What makes it plausible that a TP will still play a role in future theories is the fact that its importance for empirical success has been tested by extensive and repeated deployment. It is therefore clear that there are two ideas nested in the deployment condition. One is the idea, captured by Psillos’s criterion, that significant TPs must play a fundamental role in success in order to distinguish them from idle hypotheses; the other is the idea that the deployment of a TP must ensure that its success is not accidental. These are two distinct ideas. It might happen, for example, that a TP plays an essential role in deriving a prediction in virtue of fortuitous factors cancellation or other favorable circumstances. So, while an *intensive deployment* ensure the *fundamentality* of a TP, an *extensive deployment* founds its *robustness*. Both fundamentality and robustness are ways to articulate the complex relation between a

TP and the rest of the theory, or at least some parts of the theory (more on this in a bit). Further, while fundamentality is an atemporal articulation of this relation,² robustness concerns precisely the temporal dimension of the deployment condition that escaped Lyons's analysis: robustness, as we shall see below, is achieved over time.

In order to clarify the distinction between fundamentality and robustness, I introduce the notion of *symbolic practices*. By symbolic practices I mean all the methods customarily used in science to manipulate symbols.³ These include, but are not limited to, mathematical methods, formal tools, approximations procedures, models, heuristics, solution tricks, and any sort of way by which one can transform a symbolic expression into another symbolic expression. Symbolic practices are the set of methods adopted by a theory to "put to work" a certain TP or, in other words, to deploy it in order to set problems and to interpret solutions. By using the concept of symbolic practices, one can reformulate the two ideas of the deployment condition in the following way:

² Of course the fundamentality of a TP can change over time because it can become more or less fundamentally used. However, the relation in itself does not concern this change.

³ My discussion is especially tailored on the case of mathematical physics. I do not exclude, however, that it can be suitably extended to other branches of science by taking an appropriately enlarged notion of symbolic practices.

(F) **Fundamentality:** A TP must be *embedded* in a set of symbolic practices that lead to empirical success.

(R) **Robustness:** The symbolic practices adopted to deploy the TP must be *reliable*.

Let us begin with (F). This idea hinges on the “embeddedness” of a TP into a set of symbolic practices. An empirical success, a successful prediction or an explanation, is obtained by starting with one TP—or, better, its symbolic codification—and by deriving from it the phenomena to be treated by means of suitable manipulations. In their analysis of the path from TP to success, philosophers usually disregard the epistemic role played by symbolic manipulations of TPs. But if we neglect this important factor of the process of predicting/explaining, we are left with no other option than characterizing fundamentality as a relation between TPs, i.e., a ‘Psillosian’ criterion and then a ‘Lyonsnesque’ argument can easily prove that this falls short of providing a satisfactory notion of fundamentality. In my proposal, fundamentality is rather a relation between TP and the symbolic practices adopted to transform and manipulate it. Although intuitively clear enough, the concept of embeddedness admittedly needs further philosophical analysis. In Section 4, I provide a historical example to clarify what it means for a TP to be embedded into a set of symbolic practices.

Before discussing the example, however, I need to analyze briefly the idea of robustness. Condition (R) states that reliability, and hence robustness, is a property of the symbolic practices themselves. In other words, and this is the central point, a TP

can be made more robust by means of *historically and rationally describable strategies* conceived to enhance the reliability of symbolic practices adopted to put it to work. One way to appreciate this point is to notice that the concept of reliability has three main components. First, there is an *empirical component*, that is its connection with success. It is expected that reliable symbolic practices have led and will lead to empirical success. This is unsurprising, because it is still part of the relation between (D) and the NMA. Second, there is a *conceptual component*: reliable symbolic practices allow us to distinguish between real facts of nature and artifacts. This is the component that accounts for the non-accidentality of success and it depends on the adoption of strategies to enhance reliability. Applying symbolic practices to multiple cases, relating them with other, better understood, sets of practices (e.g., by showing structure similarities), generalizing solution methods, simplifying computation procedures, introducing redundant check routines, improving the symbolic notation, multiplying proof procedures are just a few examples of strategies used to ensure that the result of symbolic manipulation is a real information and not an artifact generated by the practice itself.⁴ Finally, there is a *historical component*. As I said above, deployment is a process extended over time. When are we justified to consider a result as reliable? This is an agent- and a context-dependent component of reliability.

⁴ This component of the concept of reliability is closely connected with the usual notion of robustness (see, e.g., (Soler et al. 2012) for an overview). Indeed, robustness has to do with the multiplications of methods of check and control as a way to distinguish what is real and what is fabricated by practices.

I submit that this component can be clarified in terms of *control*. We develop theories because we need to manipulate symbols in order to make predictions and explanations. It is reasonable to state that an agent considers reliable a theory when she has control on it, when she knows how to do things, where the theory can be applied, to what extent, what kind of information she can obtain, what kind of epistemic risks are involved in it, how to improve progressively the performance and a lot of other things related to the general idea of knowing what is going on. Thus, reliability can change over time in virtue of new information and further inquiry. This component accounts for the fact that science is an ongoing human endeavor.

To sum up, I propose to extend SelRealism in the following way:

(SelRealism+) We are entitled to consider the TP H as true or approximately true at time t if and only if:

1. H is embedded into a set of symbolic practices S
2. S is reliable
3. H and S lead to significant success

This is a more selective version of SelRealism, because the philosophical and historiographical program stemming from it extends the inquiry to the strategies adopted to improve the reliability of symbolic practices and the contingent conditions for control. As stated in condition 3, the units of analysis of SelRealism+ are TPs-*cum-*

practices rather than TPs only. In the following section, I provide an example of what I mean by intensive and extensive deployment.

4. The Coming of Age of Perturbation Theory

The *Principia Mathematica* are a supreme example of how to embed a TP, in this case the gravitational law, into a set of symbolic practices.⁵ However, Newton's mainly geometrical methods were fantastically complicated and notoriously difficult to master. A significant breakthrough in what came to be called celestial mechanics happened in the mid-1740s, when Leonhard Euler laid down the foundations of analytical perturbation theory. Euler made a number of decisive steps forward. First, he used the gravitational law to formulate general equations of motion for celestial problems. Second, he introduced the use of trigonometric series to construct approximate solutions. The use of these series also depended crucially on the gravitational law, because it satisfied the assumption that planetary orbits, even under perturbations, can be represented by a combination of periodic functions. Finally he introduced manipulation practices such as the method of the variation of

⁵ In what follows, I consider perturbation theory as the set of practices conceived to put to work the gravitational law. It must be noted that other TPs were involved (e.g., Newton's laws of dynamics) and that the gravitational law can be decomposed in further assumptions such as the action-at-a-distance, the instantaneous propagation and so forth. These considerations affect the level of detail of my example, but not the structure of my argument.

constants and the method of successive approximations to solve the equations of motion. Perturbation theory is therefore a clear example of a set of symbolic practices conceived to cast a TP into a manipulable form and to applied it to specific problems.

For the purpose of this paper, I distinguish two phases in the early history of perturbation theory. The first phase goes roughly from the mid-1740s to the mid-1760s and it concerns the cause of numerous astronomical anomalies. Newton had left behind a few conundrums that even his genius was unable to unravel. The most conspicuous of these problems was the precession of the Lunar apogee.

Newton's Lunar theory, elaborated in Book I and III of the *Principia* only managed to obtain half of the observed value. In the 1740s, there were two approaches to the issue of the Lunar apogee. The analytical approach adopted the gravitational law, or a slightly modified form of it, and tried to calculate the observed precession by analytical methods only. The physical approach supposed that the observed anomalies could be due to material causes such as a resisting medium or interplanetary vortices. It is important to realize that these approaches were compatible. Euler himself supported both the resisting medium hypothesis and the analytical approach and occasionally also proposed the use of vortices (letter to Clairaut, 30 September 1747). For several years, the best mathematicians of Europe struggled with the riddle of the Lunar apogee (Bodenmann 2010) until, on 21 January 1749, Alexis Clairaut showed that if one pushes the approximation to the second order of the perturbation, some terms that are negligible at the first order become sizable and generate the missing half of the precession (Clairaut 1752).

Clairaut's success was surely an impressive breakthrough, but what made it so impactful was not the brute fact that gravitational law had eventually led to a successful explanation. Physical hypotheses such as vortices and resisting medium also provided an explanation of the observed precession. The crucial difference lies in the fact that the gravitational law could be fully integrated with the analytical practices and then manipulated to provide suitable symbolic expressions of the precession of the apogee. That did not happen with the physical hypotheses, although not for lack of trying. Euler, for instance, tried hard to integrate the hypothesis of the resisting medium in perturbation theory, but the ensuing equations of motion were simply unmanageable (Euler 1747). Clairaut's success is eminently a story of intensive use of the gravitational law: he managed to integrate it with a set of symbolic practices and to accommodate effectively the observations.

Clairaut's feat did not close the debate on the gravitational law, though. His calculations used many case-based assumptions, simplifications, and shortcuts and its straightforward extension to more complex cases, such as the behavior of Jupiter and Saturn, was doubtful to say the least. But there was also a deeper problem. At some point in his analysis, Clairaut obtained an "arc of circle", i.e., a trigonometric function multiplied by time. Such terms are obviously unbounded and hence make the whole trigonometric series diverge. Clairaut got rid of it by ad-hoc assumptions, but the status of these unbounded terms remained unclear: they could represent an artifact of the theory, a limitation of its predictive power or even a dynamical instability of the system.

Soon, the problem of the arcs of circle become more troublesome. Euler found the same terms in his analysis of the motion of Jupiter and Saturn and in 1766 Lagrange proved that they are actually a necessary consequence of the method of successive approximations applied to astronomical problems (Lagrange 1766). Thus, in the mid-1760s, perturbation theory appeared to be a fragile set of practices which had scored some important success, but was still marred with problems of unreliability under certain conditions. From the late 1760s onwards, the issue of improving the robustness of perturbation theory became a central preoccupation of the leading mathematicians interested in physical astronomy.

There were two programs inspired by this issue. On the one hand, Lagrange tried to improve the reliability of perturbation methods *as a mathematical theory*. He carried out this project by means of multiple strategies: (1) enhancing the relation between perturbation theory and other branches of mathematics (e.g., potential theory); (2) elaborating arguments to extract information from the equations of motion without solving them (e.g., by using integrals of motion); (3) improving methods to simplify the solution procedure (e.g., Lagrange's coordinates); (4) introducing new symbolic codifications to manipulate the equations of motion (e.g., the perturbing function); (5) making the notation less cumbersome (Lagrange's coefficients). Around the same years, Laplace was also working to improve the reliability of perturbation theory, but his program adopted a different approach. He concentrated on methods to make perturbation theory a more reliable *problem-solving tool*. He developed his own method to eliminate the arcs of circle—which was based on the recalculation of the

integration constants—he imported probability theory and the equations of condition to deal with astronomical observations and devised several strategies to identify in concrete cases those elements of the equations of motion that were likely to produce sizable perturbation terms at higher order. Both Lagrange’s and Laplace’s programs scored their own successes. In the early 1780s, Lagrange proved a very general result of stability according to which the three more important orbital elements (mean motion, eccentricity, and inclination) are invariable or bounded (Lagrange 1781). Laplace, on his part, explained the decades-long problems of the anomaly in the motion of Jupiter and Saturn as well as the secular acceleration of the Moon (Laplace 1785, 1787; Wilson 1985).

5. Conclusions

In several places, Kyle Stanford has argued that any selection of enduring TPs is ultimately ungrounded and, consequently, the entire SelRealism program is unviable (Stanford 2003, 2006). In his view, there are two possible ways to select essential TPs. The first way is to trust scientists when they say that a certain posit is fundamental. However, neither commonsense, nor, more importantly, historical records support the hypothesis that scientists’ take on this matter is or should be particularly reliable. The other option is to wait and see: when a theory is superseded, one can check which TPs have survived. The reason why a selective realist cannot go with this option, however, has been summarized effectively by Peter Vickers:

If we cannot identify the working posits of a theory until it has been superseded by some other theory, then realism is no longer about identifying what we ought to believe to be true: one is always waiting for the next theory to come along to tell us which parts of our current theory are working posits. (Vickers 2013, 207)

From this, Stanford concludes that SelRealism without prospectively applicable selectivity criteria is empty and should be replaced by a more modest form of realism. But Stanford's wait-and-see stance is neither necessary nor sufficient to do the job it is supposed to do, i.e., to pick out essential TPs. It is not sufficient because there is no guarantee that the TPs survived one theory change will survive the next ones. It is not necessary because we do not need the next theory to form reasonable judgements about essential TPs. As I have shown above, science provides a variety of strategies to improve the reliability of the TP-*cum*-practices and hence good reasons to believe, *within the actual theory*, that a certain TP intensively and extensively deployed is in fact essential.

From this perspective, Stanford's argument simply sets the epistemic bar too high. By stating that the essentiality of a TP can be adjudicated only from the vantage point of the superseding theory, he implicitly challenges the realist to provide a "superselection rule" able to capture the whole history of science, a task that the realist is neither willing, nor actually requested to accomplish. By contrast, the historical and philosophical program of SelRealism+ moves from the conviction that TPs and symbolic practices follow a dynamics able to filter out inessential

components. Consequently, SelRealism+ is committed to historically identify and philosophically analyze this dynamics and to trace the genealogy of our theories in terms of the processes of codification, manipulation, and stabilization of TPs.

Ultimately, this program aims at producing new and interesting historical narratives of theory change. It remains true that the strategies making up the theoretical dynamics only provide good reasons to allocate the realist commitment. It might happen that the judgement on the reliability of the TPs-*cum*-practices change over time in virtue of further inquiry or new information. This fact, as stated above, follows from the fallibility of science as a human endeavor and, as such, should not trouble the realist.

Acknowledgements

The research for this paper has been supported by the Marie Skłodowska-Curie Actions, grant no. PIOF-GA-2013-623436.

References

Bodenmann, Siegfried. 2010. "The 18th Century Battle over Lunar Motion." *Physics Today* no. 63:27-32.

Chakravartty, Anja. *Scientific Realism* 2011 [cited 4 February 2015. Available from <http://plato.stanford.edu/entries/scientific-realism/>].

Clairaut, Alexis. 1752. "De l'orbite de la lune, en ne negligant pas les quarrés des quantités de meme ordre que les forces perturbatrices." *Memoire de L'Academie Royale des Sciences*:421-440.

Euler, Leonhard. 1747. "Recherches sur le mouvement des corps cèlestes en général." In *Opera Omnia*, 1-44. Leipzig: Teubner.

Harker, David. 2008. "On the Predilections for Predictions." *British Journal for the Philosophy of Science* no. 59:429-453.

———. 2013. "How To Split a Theory: Defending Selective Realism and Convergence without Proximity." *British Journal for the Philosophy of Science* no. 64:79-106.

Kitcher, Philip. 1993. *The Advancement of Science*. Oxford: Oxford University Press.

Lagrange, Joseph Louis. 1766. "Solution de différents problèmes de calcul intégral." In *Œuvres de Lagrange*, edited by Jean A. Serret, 609-668. Paris: Gauthier-Villars.

———. 1781. "Théorie des variations périodiques (Première partie contentant les formules générales de ces variations)." In *Œuvres de Lagrange*, edited by Jean A. Serret, 347-377. Paris: Gauthier-Villars.

Laplace, Pierre S. 1785. "Théorie de Jupiter et de Saturne." In *Œuvres de Laplace*, 95-239. Paris: Gauthier-Villars.

———. 1787. "Memoire sur les Variations seculaires des Orbites des Planetes." In *Œuvres de Laplace*, 295-306. Paris: Gauthier-Villars.

Laudan, Larry. 1981. "A Confutation of Convergent Realism." *Philosophy of Science* no. 48:19-49.

Lyons, Timothy D. 2006. "Scientific Realism and the Stratagema de Divide et Impera." *British Journal for the Philosophy of Science* no. 57:537-560.

Musgrave, Alan. 1988. "The Ultimate Argument for Scientific Realism." In *Relativism and Realism in Science*, edited by Robert Nola, 229-252. Dordrecht: Kluwer.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Soler, Lena, Emiliano Trizio, Thomas Nickles, and William C. Wimsatt. 2012. *Characterizing the Robustness of Science, Boston Studies in the Philosophy of Science*. Dordrecht: Springer.

Stanford, P. Kyle. 2003. "No Refuge for Realism: Selective Confirmation and the History of Science." *Philosophy of Science* no. 70 (913-925).

———. 2006. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press.

Vickers, Peter. 2013. "A Confrontation of Convergent Realism." *Philosophy of Science* no. 80:189-211.

Wilson, Curtis A. 1985. "The Great Inequality of Jupiter and Saturn: from Kepler to Laplace." *Archive for History of Exact Sciences* no. 33:15-290.

Worrall, John. 1989. "Structural Realism: The Best of Both Worlds?" In *Philosophy of Science*, edited by David Papineau, 139-165. Oxford: Oxford University Press.

Duhemian good sense and agent reliabilism

Famously, according to Duhem a hypothesis can never be experimentally tested in isolation, but only along with the entire theoretical scaffolding it comes with. So in the face of disagreement between theory and experiment, it is impossible to point out which hypotheses in the theory are flawed. A big question for Duhem was, how does the physicist act in such a situation of underdetermination? Which hypotheses does s/he discard, and which one(s) does s/he retain? Duhem's response was that the physicist possesses an intuitive "good sense" that directs this choice. Although good sense does not provide a rigorous, rule-based template for theory choice¹, it allows scientists to weigh evidence and be "fair and impartial" (Duhem, 218) in theory choice.

Recently, there has been much interest in drawing parallels between Duhem's good sense and ideas in virtue epistemology (VE). VE emerged in the 1980s as an approach to epistemology based on virtue ethics. In the words of Greco (2004): "Just as virtue theories in ethics try to understand the normative properties of actions in terms of the normative properties of moral agents, virtue epistemology tries to understand the normative properties of beliefs in terms of the normative properties of cognitive agents." A virtue epistemological reading of good sense as first advanced by David Stump (2007) is based on the idea that Duhem too emphasized the normative properties of the scientist qua cognitive agent and took them as a basis for legitimate scientific

¹ While "theory choice" today is generally understood in the context of contrastive underdetermination, Duhem was primarily concerned with the holist variety of underdetermination and advanced good sense in the context of the latter. But for the purposes of this paper the distinction will not matter, and I shall use "theory choice" to refer to underdetermination in general, as do all the authors I reference.

knowledge in the face of underdetermination of theory by evidence. Stump finds striking similarities particularly between Duhemian good sense and Linda Zagzebski's (1996) views of VE. Here, I discuss the views of Stump, Milena Ivanova (2010), and Abrol Fairweather (2012) in this regard and ultimately propose my own view in response which is an agent-reliabilist reading of Duhem's good sense.

Stump argues that Duhem conceived of good sense in a way that can today be understood as virtue theoretic. In particular, Stump finds similarities between good sense and ideas of VE put forward by Zagzebski (1996). As Stump tells us, Zagzebski argued that justified belief comes from a "cluster of intellectual virtues in the same way that the rightness of an act can be defined in terms of moral virtue in ethical theory" (Stump, 151). Stump argues that Duhem's good sense nicely fits in with these ideas. Good sense depends on the scientist, the cognitive agent, being "virtuous": s/he has to be, in the words of Duhem quoting Claude Bernard, a "faithful and impartial judge". Stump further provides another illuminating quote from Duhem from his lectures on German science:

"In the realm of every science, but more particularly in the realm of history, the pursuit of the truth not only requires intellectual abilities, but also calls for moral qualities: rectitude, probity, detachment from all interest and all passions. (Duhem, 1991b, p. 43)" (Stump, p. 152).

Stump notes that some of the epistemic virtues put forward by Zagzebski include intellectual sobriety, impartiality and intellectual courage and the list fits very well with Duhem's. Yet another striking similarity between Zagzebski and Duhem according to Stump is that they both appeal to non-rule-governed epistemology. Zagzebski, in making a case for an

epistemology based on ethics, says, “The idea is that there can be no complete set of rules sufficient for giving a determinate answer to the question of what an agent should do in every situation of moral choice.” (Stump, 152) Similarly, Duhem arrives at the idea of good sense when the rule-based epistemology of the physical method (i.e. strict agreement between theory and experiment) fails. As Stump says,

“Holism threatens to make testing impossible, yet Duhem believes that scientific consensus will emerge. While the pure logic of the testing situation leaves theory choice open, good sense does not. Duhem claims that the history of science shows that while there is controversy in science, there is also closure of scientific debates.” (Stump, 155)

Milena Ivanova (2010) has argued in response to Stump, that the latter is mistaken in drawing such close parallels between VE and Duhem’s good sense. She raises two main objections: first, while VE is concerned with getting to the *truth* via epistemic virtues, for Duhem, physical theory only asymptotically approaches truth – truth here being the truth of a natural order, of the “real affinities” among things. Ivanova makes this point keeping in mind Duhem’s view of a ‘perfect theory’ and the convergent nature of his realism: for Duhem, the aim of physical theory was to classify experimental laws, and a physical theory – one picked out by good sense in the face of underdetermination – constantly approached but never reached, a perfect theory which classified laws and their phenomena in exactly the way underlying metaphysical realities are really classified in nature. So her point is that while VE is concerned with getting to the truth, good sense doesn’t help us with that. But as Ivanova herself points out,

“Still, in response to this objection one can adopt the weaker thesis that even though natural

classification may not reveal the truth about the unobservable, it will be true for the observable phenomena. Also, one may argue that it is legitimate to aim at a particular epistemic goal independently of whether this goal is achievable or not.” (62)

I take her point here to be that both VE and good sense are after all in the business of truth-seeking even though attaining the truth may be impossible for with the latter.

Ivanova’s more forceful objection has to do with epistemic justification. According to her whereas VE takes epistemic virtues to be *justifications* for beliefs, Duhem did not invoke the concept of good sense to *justify* belief in one theory over another. (To reiterate, Duhem did not have a full-blown metaphysical notion of truth of a theory – but worked with the surrogate idea of truth, that a right theory approaches a transcendental, natural classification.) Rather, she argues, good sense for Duhem was more a post hoc *explanation* of the physicist’s choice: it explains the repeated success of theories at making novel predictions. According to Ivanova, what really justified belief in a theory for Duhem – i.e. the belief that it was approaching a natural classification – was the success of the theory in making correct novel predictions: She says that for Duhem, “[a scientist] is justified in believing that a theory is a natural classification only when some empirical evidence supports it or when the theory has become a ‘prophet for us’ (Duhem, 27), that is, when it has managed to make novel predictions.” (Ivanova, 62). Here’s Ivanova’s argument broken down:

- Physical theory is a classification of laws.
- In a situation where we have a theory that contradicts experimental data and are left without any means within physics to decide what to do - whether to tweak parts of the theory to accommodate the available experimental data – and if so, which parts to tweak

– or to abandon it for another theory. Somehow in the end, the scientist decides which way to go.

- The “highest test” for physical theory is to ask it to make new and novel experimental predictions.
- When the theory succeeds it is justified – in that it is taken to approach a natural classification.
- Repeatedly, the scientist sees her/his choices made in the difficult situation of underdetermination emerging successful in such predictions.
- How does this happen? There must be some innate ability or virtue in the scientist that enables him to do this: good sense.

Thus according to Ivanova, good sense is an explanation of theory choice rather than a justification for it. Moreover, according to her, Duhem doesn't say anything about good sense as a method of science: he doesn't tell us *how* exactly it directs our choice. His account of how good sense comes about and works to direct theory choice is quite thin. For Ivanova, this further shows that Duhem did not introduce it as a justification but only as a post hoc explanation.

Abrol Fairweather (2012) has argued against Ivanova's above objection and has attempted a position on Duhemian good sense that is a hybrid of Ivanova's and Stump's views. Fairweather claims to draw upon an agent reliabilist VE to do this. Reliabilism in Alvin Goldman's words, "... as a distinctive approach to knowledge is restricted to theories that involve truth-promoting factors above and beyond the truth of the target proposition." (Goldman, 2011) Fairweather's argument is that good sense results in a *reliable* process. Since Duhem's

claim is that good sense has a great “track record” and always picks out a successful theory – i.e. a theory which inevitably *correctly* makes a novel prediction – good sense produces knowledge (which here in the Duhemian context, consists in taking a predictively successful theory to be approaching a natural classification) by a *reliable* process. Good sense is a ‘truth-promoting factor’ regardless of whether the theory it picks out ultimately succeeds in novel prediction or not. It is “tracking evidentially important features of theories” (Fairweather, 10) Fairweather claims that “If a belief P is the product of a reliable capacity or process this fact constitutes evidence in favor of P.” This implies, “If the products of good sense reliably turn out to be supported by compelling new evidence, then being the product of good sense will be evidence for any theory with such a distinguished etiology.” (Fairweather, 10) So, Fairweather says, it seems that “future evidence is not required to evidentially distinguish the theory chosen by good sense, because the reliability of good sense is itself evidence supporting that theory.”

(Fairweather, 10) While I agree that agent reliabilism is the best way to understand good sense, Fairweather does not seem to give an accurate interpretation of this reading. Although he claims to provide an agent reliabilist reading of good sense, he grounds the reliability of good sense in its track record and not in its own nature or the mind where it is born. This is antithetical to agent reliabilist VE which situates reliability in the cognitive character of the agent. So it seems that Fairweather’s characterization is more along the lines of process reliabilism or simple reliabilism – according to which a belief is justified just in case it is formed via reliable processes – rather than agent reliabilism, and hence contrary to what he set out to do. His argument does not help situate good sense back into VE. Let us now turn to agent reliabilism in detail.

Greco and Agent Reliabilism: A Short Detour

As above, simple reliabilism is the view that a belief is justified just in case it is formed via reliable processes. Here the proportion of true beliefs the process results in, over time, measures reliability. Greco (1999) argues that simple reliabilism is insufficient for two reasons:

1. An agent might form a belief via fleeting or strange processes: Greco starts by noting that “Reliabilism must somehow restrict the kind of reliable process that is able to ground knowledge, so as to rule out processes that are strange or fleeting.” (Greco, 286) As an example of such processes, Greco discusses Platinga’s “The case of the epistemically serendipitous lesion” where an agent has a rare kind of a brain lesion, one that makes her believe that she has a brain lesion. There is no evidence for the lesion: there no symptoms, no testimony etc.; in fact there might even be a lot of evidence *against* it. But the agent is unable to take account of this (lack of) evidence due to the lesion. The relevant cognitive process here must no doubt be deemed very reliable, but we would not want to take the resulting belief as justified.
2. Process reliabilism doesn’t guarantee that the agent has a subjective justification of her belief. Greco says,

“[there] is a powerful intuition that knowledge does require that the knower have some kind of sensitivity to the reliability of her evidence. Sometimes this intuition is expressed by insisting that knowledge requires subjective justification. It is not enough that one's belief is formed in a way that is objectively reliable; one's belief must be formed in a way that is subjectively appropriate as well.” (285)

Greco’s solution to the above problems is agent reliabilism. According to agent reliabilism, reliability is shifted from the belief-forming process to the qualities of the agent’s

mind:

“Relevant to present purposes is Sosa's suggestion for a restriction on reliable cognitive processes; it is those processes that have their bases in the stable and successful dispositions of the believer that are relevant for knowledge and justification. Just as the moral rightness of an action can be understood in terms of the stable dispositions or character of the moral agent, the epistemic rightness of a belief can be understood in terms of the intellectual character of the cognizer.” (Greco, 287)

Following Sosa's views, Greco proposes that “knowledge and justified belief are grounded in stable and reliable cognitive character.”(Greco, 287) Accordingly, “We may now explicitly revise simple reliabilism as follows: A belief *p* has positive epistemic status for a person *S* just in case *S*'s believing *p* results from stable and reliable dispositions that make up *S*'s cognitive character.” (Greco, 287) Hence we see that reliability now has little to do with the truth of the resultant belief(s) but rather with the cognitive character of the agent.

Greco proceeds to show how agent reliabilism also solves the problem of subjective justification:

VJ: “A belief *p* is subjectively justified for a person *S* (in the sense relevant for having knowledge) if and only if *S*'s believing *p* is grounded in the cognitive dispositions that *S* manifests when *S* is thinking conscientiously.” (289)

By “thinking conscientiously”, Greco clarifies that he does not mean thinking with the purpose of finding truth, but rather the “usual state that people are in as a kind of a default mode – the state of trying to form beliefs accurately.” Greco contrasts this with epistemic “vices” such as trying to comfort oneself or trying to seek attention. Lastly, Greco points out that agent reliabilism reverses the “usual direction of analysis between virtuous character and justified

belief". While non virtue theoretic epistemologies understand virtues in terms of justified belief, here justified belief is being cashed out in terms of virtues of the cognizer. "Virtuous belief is associated with the dispositions a person manifests when she is sincerely trying to believe what is true", and "The dispositions that a person manifests when she is thinking conscientiously are stable properties of her character, and are therefore in an important sense hers." (Greco, 290) Therefore, a belief formed this way will be subjectively appropriate.

Back to Duhem

Duhem's views seem to exhibit all the features of agent reliabilism discussed above. In addition to the features of good sense and the physicist qua cognitive agent discussed so far I want to draw the reader's attention to Duhem's characterization of the different kinds of minds. For Duhem, the "strong and the narrow" mind is one capable of ordering and organizing laws and hypotheses into theories, and the "supple" mind or the "mind with finesse" – one capable of grasping a wide range of objects and at the same time able to group them logically – is the mind that produces good sense. This certainly seems to talk of "stable dispositions" in Greco's sense of the term, that reflect the "cognitive character" of the scientist. Duhem takes pains to carefully describe the mind of the physicist and discuss beliefs and attitudes *in terms of* cognitive character traits and not the other way round. i.e. Duhem talks of legitimacy of beliefs in terms of cognitive character traits; he does not talk of the traits or "epistemic virtues" so to speak, in terms of the validity of beliefs. For instance, he says about those not interested in seeing a unified system of classification erected, "Only those who affect a hatred of intellectual strength were mistaken to the extent of taking the scaffolding for a completed building." (Duhem, 103) There are several such instances where Duhem turns traditional non virtue-theoretic epistemology on its head and makes cognitive character traits basic. Now it remains to be seen if we can defend a view of

justification from good sense that goes with Greco's account. If we are successful in this, Ivanova's position will be untenable. Before going there though, let us return to Fairweather for a moment.

In addition to the argument from reliabilism, Fairweather advances another argument against Ivanova's "deflation of good sense": the position that good sense does not lend any epistemic strength or any justification to the chosen theory. The argument is that if good sense were indeed merely explanatory and post hoc as Ivanova claims, and not justificatory, then we are free to imagine a case where good sense doesn't intervene at all. After all, if good sense explains theory choice and there is no choice being made – i.e. no explanandum – we don't need an explanation. So let us suppose that we don't make any choice and just wait for a future novel prediction to make a choice and justify it. This might not be the most efficient way to choose a theory, but let us assume we do this nevertheless – for according to Fairweather, Ivanova's objection should imply the possibility of this solution. Fairweather rightly points out that in this situation we *might* again end up with an underdetermination: what if all competing theories pass the novel prediction test? Therefore, Fairweather argues, good sense must play an important epistemic role above mere explanation, in the face of such a "second level" underdetermination. But he goes further than that and says that without it, we *would* never end up with a determinate choice, even with new confirming evidence. What Fairweather is ignoring here is that future evidence *could* pick out a theory, however small the probability. It is possible that when all the options resulting from underdetermination are asked to make a novel prediction, only one succeeds, hence obviating the need for any further theory revision. But the important point is that good sense enters the scene even before such an attempt to single out a theory based on novel prediction. So the merit of good sense in my view does not lie in the inability of novel

predictions to single out a theory. It is more fundamental than that. But reasons for meriting good sense apart, let us again look at Fairweather's take on *what* the merit of good sense is.

According to Fairweather, good sense confers *uniqueness* to a theory (which, according to him, no future evidence can confer). But after good sense has uniquely picked out a theory, it is a successful novel prediction that counts as evidence in favor of the chosen theory. Fairweather makes the following interesting observation that follows from such a reading of good sense:

“This shows an interesting fact that new evidence in favor of a theory gives it a different epistemic standing depending on whether we are considering it alongside or independent of meaningful rivals. In the former case, new confirming evidence does not make a theory the determinate choice with fundamental epistemic standing. In the latter case, that same evidence determines theory choice and confers fundamental epistemic standing.” (Fairweather, 13)

So there are two “epistemic values and epistemic standings”: uniqueness, which comes from good sense, and clinching evidential support from a successful novel prediction. This way, good sense alone does not confer “fundamental epistemic standing”, and evidence alone cannot confer uniqueness. This account which recognizes an important epistemic role for both good sense and new evidence, Fairweather calls the “hybrid reading”.

My own view is that while Fairweather is right in that good sense plays a key epistemic role unlike what Ivanova says, we can go back full circle to Stump and have a proper virtue epistemological – specifically agent reliabilist – reading of good sense. I contend that good sense confers not just uniqueness, but actually does determine theory choice, also providing (an agent-

reliabilist) justification. Good sense doesn't simply pick one and put the rest "out of the running". It is not just something that prevents the proliferation of acceptable theories obtained by tweaking different parts of theories that don't agree with future experiment. Good sense provides a *basis* for the uniqueness. Just as with the problem of coming up with a realist interpretation of Duhem, this problem of the epistemic role of good sense is not easy either given the sometimes confusing nature of Duhem's claims. Nonetheless, I still think an agent-reliabilist VE reading of Duhem is possible and that Ivanova and Fairweather are mistaken.

Ivanova claims that good sense is only offered as a post hoc explanation of theory choice during underdetermination and not as a justification. I argue to the contrary. Ivanova's claim seems to be based on a purely externalist notion of justification. It seems to assume that there is one single concept of justification – specifically, externalist, evidential – and that good sense doesn't fit with it. But justification can be of many kinds. Duhem says we can "very properly decide" (Duhem, 217) between multiple theory choices using good sense. Further, he says good sense strongly "comes out in favor of" one of the choices – again implying that we are compelled to accept its judgment *even before* future experiment can ratify the choice. He goes on to say, "Pure logic is not the only rule for our judgments; certain opinions which do not fall under the hammer of contradiction are in any case perfectly unreasonable." (Duhem, 217) How do we understand such language? If an epistemic choice is proper, forceful, and reasonable, I don't see any reason we cannot properly construe it as being justified, in an *internalist* sense.

Further, Duhem does *not* introduce good sense as a merely post hoc explanation. He says, we can "properly decide" between the various options of theories using good sense. "Properly

decide” very much implies an active role for good sense *during* underdetermination. Duhem presents elaborate and careful characterizations of different kinds of minds and puts forward quite clearly, *normative* merits of cultivating/ possessing one kind of mind over the other as far as physics goes (the supple or the strong and narrow over the ample, broad and weak). Good sense is but a feature of the supple mind. It is not introduced all of a sudden as a new idea to just “save the (meta)phenomenon” of theory choice during underdetermination. It is a smooth and natural continuation of Duhem’s views on the mind of the theorist, which he articulates way before he comes to this problem of underdetermination, in one of the early chapters in *Aim and Structure*. In fact, Duhem’s view that physicists don’t actually actively choose hypotheses at all, and that they “come to his mind” when his mind is ready to receive them, clearly reveals the agent reliabilist in Duhem.

Finally, Greco’s account of agent reliabilist justification seems to lend itself to Duhem very well. Reliable cognitive character *justifies* beliefs it produces and further, it is subjectively justified: Duhem’s virtuous scientist certainly “thinks conscientiously”, following Duhem’s instructions of shunning passions and interests, and so a belief, here the belief in the theory chosen, grounded in the cognitive dispositions, here good sense, he manifests when thinking like this – is subjectively justified. So we seem to have comfortably accommodated Duhem in a full-blown agent reliabilist reading.

But what about the textual evidence cited by Ivanova, which seems to say Duhem did not think good sense justified theory choice? Why does Duhem insist that despite good sense, it is a successful novel prediction that has the final word? Why does he, in the context of resolving

underdetermination say in as many words that the method of the physicist “is justified only by experiment”? I contend that throughout *Aim and Structure*, Duhem seems to have two distinct, non-intersecting epistemologies: one of physics, and one outside of physics – which we may call philosophy. Duhem was a physicist-philosopher. He frequently claims that although there are absolutely no epistemic resources *within* physics for us to believe that physical theory latches on to a natural underlying order, we are forced to believe so by various factors outside of physics, logic and reason. It is worth noting that Duhem cites Pascal as saying that we sometimes believe for ‘reasons that reason does not know’, both in the context of theories converging on to a natural classification as well as in that of good sense during underdetermination. About the former, he says: “The opinion is a legitimate one because it results from an innate feeling of ours which we cannot justify by purely logical considerations, but which we cannot stifle completely either.”

(Duhem, 102) Further:

“No language is precise enough and flexible enough to define and formulate them; and yet, the truths which this common sense reveals are so clear and so certain that we cannot either mistake them or cast doubt on them; furthermore, all scientific clarity and certainty are a reflection of the clarity and an extension of the certainty of these common-sense truths.” (Duhem, 104)

Since Duhem attributes good sense to similar patterns of thinking, we can associate his above assertions about the legitimacy of beliefs not borne out of logic, with good sense as well. Given Duhem’s commitment to the moral goodness and the intellectual acuity of the supple, strong and narrow minds, it is very unlikely that he would think that epistemic ends justify the means (here, successful novel prediction justifying that which chose the theory, i.e. good sense). Reliabilism in fact expressly turns this around and say it is the means (by virtue of their

reliability) that justify the ends. So beliefs that arise from good sense are *justified* from an (internalist, deontological) agent reliabilist perspective. The justification Duhem talks about when he says that the methods of the physicist are justified by experiment should be when we are strictly within the context of physics: there it is Duhem qua physicist speaking. But from a broader, philosophical perspective, Duhem rather means, I think, that experiment *validates* the choice and confers *certainty* on it. But we can have justification without certainty, like in agent reliabilism. In simpler terms, the *reasons* for which the physicist chooses a theory are grounded in her good sense. However, the successful novel prediction will no doubt make the choice certain.

Thus, Ivanova is mistaken in arguing that good sense does not provide justification. Fairweather's hybrid reading is inadequate as well for it ignores the justification offered by a proper agent reliabilist reading of good sense. I argue that a proper agent reliabilism accommodates Duhem as a virtue epistemologist very well and shows us that good sense does offer justification for theory choice. Importantly, I have shown that it is certainly not a post hoc explanation but a part and parcel of Duhem's overall views on the mind of the physicist.

References

- Duhem, Pierre. (1954). *The aim and structure of physical theory*. Princeton: Princeton University Press.
- Fairweather, A. (2012) 'The Epistemic Value of Good Sense' *Studies in the History and Philosophy of Science* <http://philpapers.org/archive/FAITEV.pdf>

-
- Goldman, Alvin. 'Reliabilism', *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/reliabilism/>.
 - Greco, J. 1999. 'Agent reliabilism' in *Philosophical Perspectives* 13: 273-296.
 - Ivanova, M. (2010). 'Pierre Duhem's good sense as a guide to theory choice'. *Studies in History and Philosophy of Science*, 41, 58–64.
 - Stump, David. (2007). Pierre Duhem's virtue epistemology. *Studies in History and Philosophy of Science*, 38, 149–159.

There *Is* a Special Problem of Scientific Representation

(Word count: 4998)

Abstract: Callender and Cohen (2006) argue that there is no need for a special account of the constitution of scientific representation. I argue that scientific representation is communal and therefore deeply tied to the practice in which it is embedded. The communal nature is accounted for by *licensing*, the activities of scientific practice by which scientists establish a representation. A case study of the Lotka-Volterra model reveals how the licensure is a constitutive element of the representational relationship. Thus, any account of the constitution of scientific representation must account for licensing, meaning that there *is* a special problem of scientific representation.

1. Introduction

According to many philosophers of science, representation in scientific practice is different from representation in other disciplines, like art and language. This claim is denied by Craig Callender and Jonathan Cohen (2006), who argue that representation is the same across disciplines. In this paper, I will argue that their view leaves the communal nature of scientific representation unexplained. To explain why scientific representation is dependent upon practice, I will introduce the concept of licensing, in which the targets of representational vehicles are determined through various activities performed by scientists in accord with broader scientific practice. I will argue that licensure is a constitutive feature of representation in science, indicating that there *is* a special problem of scientific representation.

2. Callender and Cohen's View

On Callender and Cohen's evaluation, much of the literature on scientific representation has been "concerned with non-issues" (2006, 67). Specifically, they think there is no reason for philosophers of science to give a special account of the "constitution question:" "What constitutes the representational relation between a model and the world?" (2006, 68). In response to this question, they make a few observations. One is that it is "economical and natural to explain some types of representation in terms of other, more basic types of representation" (2006, 70). They also identify a general desire to have a consistent account of how "entities other than models—language, pictures, mental states, and so on—...represent the very same targets that models represent" (2006, 71). For these reasons, they suggest that

“scientific representation is just one more special case of derivative representation” (2006, 75). That is to say that the representational nature of scientific vehicles is explained in the same way that the representational nature of linguistic entities, artwork, etc. is explained. In each case, and in every practice, the representational nature in question will be reduced to a more fundamental representational entity. So, e.g., the representational nature of a word, a painting, and a scientific model will each be explained in terms of the representational nature of mental states.

On Callender and Cohen’s view, representation is purely stipulative: “virtually anything can be stipulated to be a representational vehicle for the representation of virtually anything...” (2006, 74). Of course, it is not the case that *any* stipulated representation will actually be useful for scientific aims. Thus, they identify pragmatic constraints which delimit scientific representation. However, they make it quite clear that these constraints are delimiting *already-existing* representations. As such, the pragmatic constraints are not a part of an account of the constitution of representation itself: “the questions about the utility of these representational vehicles are questions about the pragmatics of things that are representational vehicles, not questions about their representational status per se” (2006, 75).

If Callender and Cohen are correct, then we are left rethinking a rather extensive literature on scientific representation which typically begins with the assumption that there *is*

something special about representation in science.¹ As one example among many, Mauricio Suárez (2004) defends an inferential conception of scientific representation. His account takes careful notice of the aims of scientific practice, noting that mere stipulation (what he calls “representational force”) is insufficient for representation in science. To be a *scientific* representation, a vehicle must also permit surrogate reasoning which “allows competent and informed agents to draw specific inferences regarding [a target]” (2004, 773). If we accept Callender and Cohen’s view, then Suárez’s account and the many others like it do nothing more than identify some of the typical pragmatic strategies employed in delimiting representations for scientific uses (Callender and Cohen 2006, 78).

3. Private Reminiscence and Communal Representation

In order to show that the extensive literature on scientific representation has not been addressing a non-issue, I will need to show that there is a special problem of scientific representation, a feature unexplained by Callender and Cohen’s account. I submit that the relevant feature in need of special explanation is the communal nature of scientific representation, that it inherently involves reference to the practice. To see why Callender and

¹ For more accounts which answer the constitution question in a distinct way, see the work of Ronald Giere (1988, 2004), Bas van Fraassen (1980, 2008), RIG Hughes (1997), Steven French, James Ladyman, and Otávio Bueno (French and Ladyman 1999; Bueno and French 2011), and Gabriele Contessa (2007). For an overview of these accounts of scientific representation among others, see Brandon Boesch (2015) and Mauricio Suárez (2015).

Cohen's view is unable to account for the communal nature of scientific representation, consider what I call 'reminiscence', a representational relationship which lacks the same communal feature. It is defined schematically as the following:²

Some X is reminiscent of some Y for some agent A provided that when A thinks about or experiences X, she thinks about or experiences Y and attributes some connection between X and Y.

So, for example, a drawing can be reminiscent of my nephew, the smell of honeysuckle can be reminiscent of golfing, etc.

There are three noteworthy features of reminiscence. First, the representational nature of reminiscence can be reduced to the representational nature of more fundamental entities. For example, I can explain the drawing's reminiscence of my nephew in virtue of the mental state produced by the drawing (which is about my nephew, who created it). Second, stipulation is sufficient to create an instance of reminiscence. For example, I could draw a symbol on my hand which I create for the sake of reminding me to buy bread from the store. The reminiscent relationship exists because of my stipulative act. Finally, any limitations of reminiscent relationships will be made for pragmatic reasons. For example, it would be for pragmatic reasons that I make the symbol on my hand look like a loaf of bread.

² I should note that the account of reminiscence here is not meant as a detailed explanation of this concept, but only as an analogy to draw a point about representation.

These three features of reminiscence are noteworthy because they are shared by Callender and Cohen's view of scientific representation. In fact, from Callender and Cohen's perspective, the only major difference between the two concepts would be the particular aims for which each relationship is utilized. While important, these different aims alone are insufficient to explain a key dissimilarity between scientific representation and reminiscence: while reminiscence can be private, scientific representation is necessarily communal. That reminiscence can be private can be seen from the fact that discussions of reminiscence can terminate in disagreement. For example, no one is ultimately 'correct' about whether or not someone is reminiscent of someone else. This is because reminiscence is agent-relative and so depends only upon some particular agent and her mental states.

Scientific representation relies on much more. As Suárez has argued, "representation is not at all 'in the mind' of any particular agent. It is rather 'in the world', and more particularly in the social world – as a prominent activity or set of activities carried out by those communities of inquirers involved in the practice of scientific modelling" (2010, 99). Scientific representation is not isolated from the practice in which it is embedded. It is necessarily communal.³ The communal nature is demonstrated from the fact that representational vehicles demonstrate autonomy from individual scientists and their mental

³ The view of representation argued for in this paper echoes many of the points made by Ludwig Wittgenstein's in his 'Private Language Argument' where he argues that meaning is necessarily communal (1953/2009, 95^e-111^e).

states.⁴ For example, a scientist's rogue stipulation that the Lotka-Volterra model (which represents predator-prey relations) represents population change due to genetic drift does not count as an instance of scientific representation. This is not only because it does not (pragmatically) allow for meaningful insights, but also because it ignores and discounts the autonomous elements of the model as understood by the broader scientific community.⁵ The autonomous elements are seen in the materiality or historicity of the representational vehicle; in its development, reception, and contemporary use. Understanding how and why the scientific object represents its target requires paying attention to these communal features. That is to say that the communal nature is partially *constitutive* of the representational relationship. Callender and Cohen's account of scientific representation does not sufficiently account for these constitutive communal elements, as will be shown more explicitly below.

4. Licensing

Explaining the communal nature of scientific representation requires that attention be given to the material, autonomous dimensions of the representational vehicle in terms of its

⁴ This point has already been made specifically with regard to models by Morrison and Morgan (1999). Here, I am extending a similar point to other representational vehicles, including things like diagrams and figures.

⁵ Of course, there may be disagreements and developments internal to the practice about how to use some representation, but these disagreements and developments are *part of the practice*.

development, reception, and use. All of these features partially establish a scientific representation, through an activity I call *licensing*. Licensing is the set of activities of scientific practice by which scientists establish the representational relationship between a vehicle and its target. It is itself a constitutive element of the representational relationship: it is a critical part in explaining how and why some vehicle represents its target. Seeing the sorts of activities involved in licensing and how they partially constitute the representational relationship will require that we pay close attention to the historical development, reception, and use of actual instances of scientific representation.

4.1 Licensing in Artistic Representation

A similar sort of licensing is present in representation in art, and so an initial pass on the concept as it applies to artistic practice will be helpful to draw an analogy to licensing in science.⁶ To see the role of licensing in artistic representation, consider an example. The mere stipulation that Pablo Picasso's *Guernica* should represent the pain of cyberbullying is clearly insufficient to make it represent this target. Understanding how *Guernica* is representational involves an awareness of communal features: Picasso's intentions within the environment in which he created the painting, how the painting was received by viewers in the years following its creation, and how it is understood today. With these features in mind,

⁶ It is somewhat contentious to draw conclusions about the nature of representation in science by appeal to art; see e.g. Bueno and French (2011). Nonetheless, it is a common technique in discussions of scientific representation; see e.g. Suárez (2004).

it is clear that *Guernica* represents the pain and suffering of the people of Guernica who had been bombed by axis forces at the request of Francisco Franco and the Spanish Nationalists. The licensing here is a constitutive element of *Guernica*'s representational nature: without these features, it is not clear whether or how the painting would manage to represent anything at all.

Licensing also occurs outside of the scope of authorial intent, when the artistic community comes to accept that a piece of art is representational in a way that was not intended by the author. A good example can be taken from an anecdote related by the author Flannery O'Connor:

[A] student asked me...: "Miss O'Connor, what is the significance of the Misfit's hat?" Of course, I had no idea the Misfit's hat was significant, but finally I managed to say, "Its significance is to cover his head." (1988, 853)

The Misfit is a key character in O'Connor's famous short story, "A Good Man is Hard to Find," and, as such, it would not be surprising for his wardrobe to be importantly representational. Her answer indicates that while she did not intend any representational target for the hat, there may yet be one. If the hat is representational, it will not be due to her authorial intent, but rather due to the views of the broader artistic community.

Let me make it very clear that the licensure so far described is not already accounted for by elements of Callender and Cohen's account. First, notice that none of these means of licensing is a mere pragmatic limitation of already existing representations. It is not as if *Guernica* represents anything and everything, but is then *limited* by the contexts of Picasso,

audiences, and art historians. These contexts are a crucial part of understanding why it represents at all. Nor is the licensing mere stipulation. O'Connor leaves it open that there may be a representational target for the Misfit's hat, even though she did not stipulate one. A single reader's stipulation alone is insufficient to make it a representation, since the target must also fit well with the Misfit's characteristics, with O'Connor's general themes as understood by literary critics and audiences alike, and so on. Once again, these contexts are a critical part of establishing the representational nature of the hat.

4.2 Licensing in Scientific Representation: A Case Study

The unique aims of science indicate that the licensing of scientific representation is of a different kind than the licensing in art. All the same, licensing similarly plays a critical role in establishing scientific representation. According to Tarja Knuuttila, case studies of scientific representation have revealed that it is "a complicated phenomenon" and "a laborious art" (2014, 304). Understanding the nature of licensing and its role in the complexities of scientific representation will be best accomplished by examining the complicated features seen in the context of a case study. Examples could be made of any type of representational vehicle, like the masterful case study of a scientific figure made by Bruno Latour (1999). I will take as my example the Lotka-Volterra model, since its development exhibits interesting features, many of which have already been widely discussed by other philosophers (e.g. Knuuttila and Loettgers 2011, forthcoming).

As mentioned above, the Lotka-Volterra model is used by ecologists to represent predator-prey relations. It had its beginnings in the independent work of two different

scientists, Vito Volterra and Alfred Lotka. In understanding the representational nature of this model, it is important to pay attention to the licensing through its historical development. This attention includes noticing things like the way that the construction of the model by Lotka, Volterra, and others has been responsive to certain theoretical and empirical aims. These historical and practice-centered features of the model's development reveal the partial autonomy of its representational nature. These features constitute the licensing which is itself partially constitutive of the representational nature of the model since understanding how and why the model represents its targets requires attending to these features. Let us now turn to examine these features in more detail.

Consider first the development of the model by Volterra, who was "motivated by the goal of reproducing the kind of oscillating behavior that was observed empirically in fishery statistics" (Knuuttila and Loettgers forthcoming, 19). His aim to address a theoretical question with an empirically useful model is central not only to understanding how the model historically came about, but in understanding how it represents its targets. Consider how Volterra described his project and the aims which permeate his description:

Let us seek to express in words the way the phenomenon proceeds roughly: afterwards let us translate these words into mathematical language. This leads to the formulation of differential equations. If then we allow ourselves to be guided by the methods of analysis we are led much farther than the language and ordinary reasoning would be able to carry us and can formulate precise mathematical laws. These do not contradict the results of observation. Rather

the most important of these seems in perfect accord with the statistical results.

(1928, 5)

Volterra's actual process of moving from words, to equation, to application of results (for both theoretical and empirical purposes) first involved creating an equation to account for the population change of a single species. He then added additional species and modelled interactions under different conditions, including, notably, contending for the same food and the predation of one species upon the other. Using these models, he demonstrated "three fundamental laws of the fluctuations of the two species living together" (1928, 20). He then applied these theoretical laws of predator-prey relations to the empirical case which had prompted his analysis, the peculiar rise in predator populations during the decrease of fishing of prey populations in the Adriatic Sea during World War I (1928, 21).

Why does Volterra's model represent these theoretical features of predator-prey relations? Why does it represent the populations of fish in the Adriatic during World War I? It represents these targets because, through a series of steps of analysis, revision, and development, each of which was responsive to certain theoretical and empirical aims understood and described in his account, Volterra *established* this representational nature. Indeed, as explained by Knuuttila and Loettgers (forthcoming), the historical development of this model has a much more extended history than the one Volterra described in the two papers where he first introduced it (1926, 1928). The model is a representation of its target not by mere stipulation and pragmatic constraint, but through careful and attentive construction of equations which ensure that the model functions in the wider theoretical

contexts and can explain the relevant empirical aims. In short, the model represents its targets because Volterra so *licensed* it by building into the model these external, autonomous representational features. Without these features, how or what would it represent?

Consider another instance of licensing in the development of the Lotka-Volterra model, this time by Lotka. His development proceeded with a different aim than Volterra: “instead of starting from the different simple cases and generalizing from them, he developed a highly abstract and general model template that could be applied in modelling various kinds of systems” (Knuuttila and Loettgers forthcoming, 13). He began by creating a very general equation which described “evolution as a process of redistribution of matter among the several components...of the system” (Knuuttila and Loettgers forthcoming, 15). In two papers (1920a, 1920b), Lotka applied this general equation to particular cases in biology and chemistry, in each case coming to theoretical conclusions about the systems in question. For example, in applying the equation to a predator-prey system, he concluded that there would be “undamped oscillation continuing indefinitely” among the two populations (1920a, 414). Lotka did not specifically apply the results to any empirical data, but instead used his results to come to theoretical conclusions about these relationships which he then connected to theoretical ecological principles drawn from Herbert Spencer’s *First Principles* (1920a, 414).

Why does Lotka’s model represent its theoretical target? What constitutes this representational relationship? Any attempt to explain the representational relationship must reference the way in which Lotka derived his general equation and the way in which he applies it to the specific cases. That is to say, the representational nature of the model is

constructed through the scientific activities performed by Lotka during the development of the model. Lotka does not merely stipulate that his model targets predator-prey relationships. Instead, he builds this ability into the model during the development of the general equation and further constructs this ability in his application of the question to specific targets. In so doing, he partially constructs the representational nature of the model—he licenses it as a representation through activities in accord with the broader practice.

The Lotka-Volterra model's history since its initial development is long and complex. As described by Alan Berryman (1992), one development was a shift in the 1940s to the use of a logistic formulation which allowed for attention to be placed on predator-prey ratios rather than products. Another development, which occurred around the same time, was the use of a predator functional response which introduced a nonlinear rate of death for the prey. These developments license new representational targets by expanding and altering the model to make it responsive to different theoretical or empirical aims, by removing idealizations, or otherwise by allowing for different theoretical conclusions. Many other variations of the Lotka-Volterra model exist, licensed by similar developments. Additionally, the original formulation of the model is still used in introductory textbooks on ecology (see, e.g. Cain, Bowman, and Hacker 2008). The representational nature of the model in each of these cases is partially established by these features of the model which stand independent of any mental states of scientists and students alike. In short, the constitution of the representational nature of the Lotka-Volterra model relies deeply upon these historical features of licensing as understood by the broader scientific community.

Let me briefly underscore the importance of these activities of licensing to the representational nature of the Lotka-Volterra model by imagining a scenario in which these features are absent. Suppose that Volterra and Lotka had proceeded differently. Suppose that they began, for no particular reason, by drawing a five-pointed star and stipulated that it represented predator-prey relations. What is the status of this star, qua representation? It is not as if the star *really* is a scientific representation of predator-prey relations albeit a bad representation (because it does a poor job of meeting certain pragmatic constraints). Rather, the star plainly fails to be a scientific representation at all. Scientific representations are constructed to assist in answering certain questions, explaining certain phenomena, understanding certain target systems. It is through licensing that scientists build into the vehicle the features capable of achieving these aims. A vehicle without licensing does not have this ability and so it is not just a bad representation. It is not a representation *at all*. Indeed, a discussion of the representational nature of vehicles which lack these features is either infelicitous or involves an equivocation of the word ‘representation.’ A view of scientific representation which equally counts both the star and the Lotka-Volterra model as full scientific representations, even if it specifies one as good and one as bad, underestimates the role of these historical features of the model. They are not external to the representational nature of the vehicle, but are themselves an essential constitutive feature of this representational nature: without these features, the vehicle is not a scientific representation at all.

5. The Special Problem of Scientific Representation

If I am right that licensing is a necessary constitutive feature of scientific representation which explains its communal nature, then contrary to Callender and Cohen's suggestion, we cannot pull the question of the constitution of representation away from questions of practice. A scientific object represents its target not (only) because there is some stipulation and pragmatic constraint, but also in virtue of licensing: the context in which it was created, the application of theoretical and empirical constraints, the awareness of and management of idealizations, and the history of its reception and use. Accounting for whether and how a scientific object represents its target will always require reference to these features which partially establish the representational nature. Thus, there *is* a special problem of scientific representation.

I should note that I am not here arguing for a stronger counter claim to Callender and Cohen which says that accounts of the representational nature of mental states are without *any* value to the constitution question of scientific representation. But my argument does indicate that an account of the representational nature of mental states *alone* is insufficient to account for scientific representation. Even if tomorrow we had a solid, universally accepted account of the representational nature of mental states, we would not yet have a complete account of scientific representation. We would still need an account of the deep reliance that it has upon the practice in which it is embedded. Thus, while our discussion of the constitution of scientific representation might include reference to the representational nature of mental states, it must also include reference to what I have described here as the licensing by the practice.

A different concern is that the use of the word ‘special’ is a bit deceptive. What I have identified here as the ‘special’ problem of scientific representation turns out to be a common feature of representation across disciplines, since, for example, I have suggested that it holds of artistic representation as well. While it is true that, according to my argument, an account of artistic representation will likely take account of licensing as well, it does not indicate that it is the *same type* of licensing in both practices. Indeed, given the unique aims that mark off scientific practice, its licensing can reasonably be expected to be correspondingly unique. That is to say that understanding, knowing, or explaining the empirical world are special aims, and therefore subject to special sorts of licensing. Scientific representation remains special because these features merit special attention.

We might also wonder whether it is right to continue to discuss scientific representation as a whole. If understanding representation in science requires in part that we understand the way in which scientists of a practice develop, utilize, and adapt these representational devices, then it is at least possible that these activities will be different within different domains. For example, the licensure of representations in physics might be rather different from that of economics. My suspicion is that, given the common broad scale aims of the various domains, we can still say some general things about representation in science as a whole. Nonetheless, we would do well to pay attention to representation as it occurs in these more localized contexts. Moving forward from this conclusion to develop further insights about the nature of scientific representation will involve analyzing specific representational objects or strategies as they occur in scientific practice, perhaps taking hints and clues from

in-the-field investigations like those conducted by sociologists of science, e.g. those in Lynch and Woolgar (1990), Latour (1999), and Coopmans et al. (2014).

6. Conclusion

Though Callender and Cohen's view remains a formidable approach to the constitution question of scientific representation, I have endeavored in this paper to show why their account is insufficient, and thus why this question merits continued attention by philosophers of science. Representation in science is deeply tied up with the practice in which it is embedded. The communal nature of scientific representation can be seen in the way that science, as a practice, partially constructs its representations through the activities of licensing. The licensing is not the pragmatic limitation of some already existing representations, but is itself a constitutive element of the representational relationship. Any account of what it is for a scientific object to represent its target will necessarily involve reference to licensing. Thus, there *is* a special problem of scientific representation.

Bibliography

- Berryman, Alan. 1992. "The Origins and Evolution of Predator-Prey Theory." *Ecology* 73: 1530-1535.
- Boesch, Brandon. 2015. "Scientific Representation." *Internet Encyclopedia of Philosophy*.
<http://www.iep.utm.edu/sci-repr/>
- Bueno, Otávio, and Steven French. 2011. "How Theories Represent." *British Journal for the Philosophy of Science* 62: 857-894
- Cain, Michael, William Bowman, and Sally Hacker. 2008. *Ecology*. Sunderland, MA: Sinauer Associates, Inc.
- Callender, Craig, and Jonathan Cohen. 2006. "There Is No Special Problem About Scientific Representation." *Theoria* 21: 67-85.
- Contessa, Gabriele. 2007. "Scientific Representation, Interpretation, and Surrogate Reasoning." *Philosophy of Science* 74: 48-68.
- Coopmans, Catelijne, Janet Vertesi, Michael E. Lynch, Steve Woolgar (eds.). 2014. *Representation in Scientific Practice Revisited*. Cambridge, MA: MIT Press.
- French, Steven and James Ladyman. 1999. "Reinflating the Semantic Approach." *International Studies in the Philosophy of Science* 13: 103-119.
- Giere, Ronald. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.
- . 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71: 742-752.

Hughes, R.I.G. 1997. "Models and Representation." *Philosophy of Science* 64 (Proceedings): S325-S336.

Knuuttila, Tarja, and Andrea Loettgers. 2011. "The Productive Tension: Mechanisms Vs. Templates in Modeling the Phenomenon." In *Models, Simulations, and Representations*, ed. P. Humphreys and C. Imbert, 3-24. New York: Routledge.

———. Forthcoming. "Modelling as Indirect Representation? The Lotka-Volterra Model Revisited." *British Journal of Philosophy of Science*, in press.

Knuuttila, Tarja. 2014. "Reflexivity, Representation, and the Possibility of Constructivist Realism." In *New Directions in the Philosophy of Science*, ed. M. C. Galavotti, S. Hartmann, M. Weber, W. Gonzalez, D. Dieks, and T. Uebel, 297-312. Dordrecht, The Netherlands: Springer.

Latour, Bruno. 1999. Circulating Reference. In *Pandora's Hope*. Cambridge: Harvard University Press.

Lotka, Alfred. 1920a. "Analytical Note on Certain Rhythmic Relations in Organic Systems." *Proceedings of the National Academy of Arts and Sciences* 42: 410-415.

———. 1920b. "Undamped Oscillations Derived from the Law of Mass Action." *Journal of the American Chemical Society* 42: 1595-1598.

Lynch, Michael E., and Steve Woolgar (eds.). 1990. *Representation in Scientific Practice*. Cambridge: MIT Press.

Morgan, Mary, and Margaret Morrison (eds.). 1999. *Models as Mediators: Perspectives on Natural and Social Science*. New York: Cambridge University Press.

O'Connor, Flannery. 1988. "The Catholic Novelist in the Protestant South." In *Flannery O'Connor: Collected Works*, 853-864. New York: Literary Classics of the United States.

Suárez, Mauricio. 2004. "An Inferential Conception of Scientific Representation."

Philosophy of Science 71: 767-779.

———. 2010. "Scientific Representation." *Philosophy Compass* 5: 91-101.

———. 2015. "Representation in Science." In *The Oxford Handbook of Philosophy of Science*, ed. P. Humphreys. New York: Oxford.

van Fraassen, Bas C. 1980. *The Scientific Image*. New York: Oxford University Press.

———. 2008. *Scientific Representation: Paradoxes of Perspective*. New York: Oxford University Press.

Volterra, Vito. 1926. "Fluctuations in the Abundance of a Species Considered Mathematically." *Nature* 128: 558-560.

———. 1928. "Variations and Fluctuations of the Number of Individuals in Animal Species Living Together." *Journal du Conseil International Pour l'Exploration de la Mer* 3:3-51.

Wittgenstein, Ludwig. 1953/2009. *Philosophical Investigations*, trans. G.E.M. Anscombe, P. M. S. Hacker, and J. Schulte. Malden, MA: Wiley-Blackwell.

Dissolving the missing heritability problem

Abstract: Heritability estimates obtained in genome-wide association studies (GWAS) are much lower than those of traditional quantitative methods. This has been called the “missing heritability problem”. By analyzing and comparing these two kinds of methods, we first show that the estimates obtained by traditional methods involve some terms that GWAS do not. Second, the estimates obtained by GWAS do not take into account epigenetic factors transmitted across generations, whilst they are included in the estimates of traditional quantitative methods. Once these two factors are taken into account, we show that the missing heritability problem can be largely dissolved. Finally, we briefly contextualize our analysis within a current discussion on how non-additive factors relate to the heritability estimates in GWAS.

1. Introduction.

One pervasive problem encountered when estimating the heritability of quantitative traits is that the estimates obtained from Genome-Wide Association Studies (GWAS) are much smaller than that calculated by traditional quantitative methods. This problem has been called the missing heritability problem (Turkheimer 2011). Take human height for example. Traditional quantitative methods deliver a heritability estimate of about 0.8, while the first estimates using GWAS were 0.05 (Maher 2008). More recent GWAS methods have revised this number and estimate the heritability of height to be at most 0.45 (Yang et al. 2010; Turkheimer 2011). Yet, half of the heritability is still missing.

In quantitative genetics, heritability is defined as the portion of phenotypic variation in a population that is caused by genetic difference (Downes 2015). Traditionally, this portion is estimated by measuring the phenotypic resemblance of genetically related individuals without identifying at the molecular level (more particularly the DNA level) the genetic causes of phenotypic variation. GWAS have been developed in order to locate the DNA sequences that influence the target trait and estimate their effects, especially for common complex diseases such as obesity, diabetes and heart disease (Visscher et al. 2012; Frazer et al. 2009). As for height, almost 300 000 common DNA variants in human populations that associate with it have been identified by GWAS (Yang et al. 2010). Granted by many that the heritability estimates obtained

by traditional quantitative methods are quite reliable, the method(s) used in GWAS have been questioned (Eichler et al. 2010).

A number of partial solutions to the missing heritability problem have been proposed, with most of them focusing on improving the methodological aspects of GWAS in order to provide a more accurate estimate (e.g., Manolio et al. 2009; Eichler et al. 2010). Some authors have also suggested that heritable epigenetic factors might account for part of the missing heritability. For instance, in Eichler et al. (2000, 488), Kong notes that “[e]pigenetic effects beyond imprinting that are sequence-independent and that might be environmentally induced but can be transmitted for one or more generations could contribute to missing heritability.” Furrow et al. (2011) also claim that “[e]pigenetic variation, inherited both directly and through shared environmental effects, may make a key contribution to the missing heritability.” Others have made the same point (e.g., McCarthy and Hirschhorn 2008; Johannes et al. 2008). Yet, in the face of this idea one might notice what appears to be a contradiction: how can *epigenetic* factors account for the missing heritability, if the heritability is about *genes*?

To answer this question as well as to analyze the missing heritability problem, we compare the assumptions underlying both heritability estimates in traditional quantitative methods and those in GWAS. We argue that a) the heritability estimates of traditional methods include some terms associated with broad-sense heritability (H^2), as opposed to narrow-sense heritability (h^2); b) although GWAS are supposed to get h^2 , h^2 relies on an evolutionary concept of the gene

that can include epigenetic factors while heritability estimates obtained from GWAS do not. With these two points being illustrated, we expect the missing heritability problem to be largely dissolved as well as setting the stage for further discussions.

The remainder of the paper will be divided into three parts. First, we briefly introduce how heritability is estimated in two traditional methods, namely twin studies and parent-offspring regression. We show that the estimates obtained by each methods include *some* non-additive elements and consequently correspond neither to H^2 nor to h^2 , but to a notion in between which we term “broader-sense heritability”. Second, we outline the basic rationale underlying GWAS and illustrate that they estimate heritability by considering solely DNA variants. By arguing that the notion of additive genetic variance does not necessarily refer to DNA sequences but can also refer to epigenetic factors in traditional quantitative methods, we show that the notion of heritability estimated in GWAS is more restrictive than that of traditional quantitative methods, and term this notion “DNA-based narrow-sense heritability”. Finally, in Section 4, based on the conclusions from Section 2 and Section 3, we claim that the gap between the heritability estimates of traditional quantitative methods and those of GWAS can be explained away in two major ways. One consists in recognizing that if non-additive variance was removed from the estimates obtained via traditional methods, they would be lower. The other consists in recognizing that if epigenetic factors were taken into account by GWAS, the heritability estimates obtained would be higher. We conclude Section 4 by showing how our analysis sheds

some light on a discussion about the role played by non-additive factors in the missing heritability problem. Because human height has been “the poster child” of the missing heritability problem (Turkheimer 2011, 232), we will use this example to illustrate each of our points.

2. Heritability in Traditional Quantitative Methods.

According to quantitative genetics, the phenotypic variance (V_P) of a population can be explained by two components, its genotypic variance (V_G) and its environmental variance (V_E).

In the absence of gene-environment interaction and correlation, we thus have:

$$V_P = V_G + V_E \quad (1)$$

From there broad-sense heritability (H^2) is defined as:

$$H^2 = \frac{V_G}{V_P} \quad (2)$$

V_G can further be portioned into the additive genetic variance (V_A), the dominance genetic variance (V_D) and the epistasis genetic variance (V_I). We have:

$$V_P = V_A + V_D + V_I + V_E \quad (3)$$

where V_A is the variance due to hypothetical genes making an equal and additive contribution to the trait studied (e.g., height). V_D is the variance due to interactions between alleles at one locus for diploid organisms, and V_I is the variance due to interactions between alleles from different loci. V_D and V_I together represent the variance due to particular combinations of genes of an organism.

Since genotypes of sexual organisms recombine at each generation via reproduction, dominance and epistasis effects are not transmitted stably across generations, only additive genetic effects are. Therefore, V_A is the variance due to stably transmitted genetic effects. Narrow-sense heritability (h^2) measures to what extent variation in phenotypes is determined by the variation in genes transmitted from parent(s) to offspring (Falconer and Mackay 1996, 123). It is defined as:

$$h^2 = \frac{V_A}{V_P} \quad (4)$$

h^2 is important in breeding studies and is used by evolutionary theorists who are interested in making evolutionary projections of a trait within a population across generations.

To know h^2 , both V_A and V_P must be known. V_P , for most quantitative traits (including height), can be directly estimated by measuring individuals. However, there is no direct way to estimate V_A in traditional quantitative methods. The traditional way to estimate it requires two elements. First, one needs a population-level measure of a phenotypic resemblance of family

relative pairs¹. This measure is obtained by calculating the *covariance* of the phenotypic values for those pairs. The choice of what sort of relatives to use depends on what data is available. The second element is the genetic relation between family pairs. It indicates the percentage of genetic materials the pairs are expected to share. With these two elements, one can estimate how much the genes shared contribute to the phenotypic resemblance. In a large population with different phenotypes, one can then estimate how much the additive genetic difference contributes to phenotypic difference in this population, which estimates h^2 .

For simplicity, traditional quantitative methods usually assume that there is neither gene-environment interaction nor correlation (Falconer and Mackay 1996, 131). Thus the covariance between the phenotypic values (e.g., height) of pairs equals to additive genetic covariance, dominant and epistasis genetic covariance, plus the environmental covariance. A general equation for traditional quantitative methods can be written as follows:

$$\begin{aligned} Cov(P_1, P_2) &= Cov(A_1 + D_1 + I_1 + E_1, A_2 + D_2 + I_2 + E_2) = \\ &Cov(A_1, A_2) + Cov(D_1, D_2) + Cov(I_1, I_2) + Cov(E_1, E_2) \end{aligned} \quad (5)$$

where indexes “1” and “2” represent the two family members for each pair studied.

$Cov(P_1, P_2)$ is the covariance between the phenotypic values of one individual with the other.

¹ Or the mean values of their class (e.g., offspring) depending on the particular method used.

A , D , I and E represent additive effects, dominant effects, epistasis effects and environmental effects respectively.

The most commonly used traditional methods for estimating heritability are twin studies. In these studies one already knows that monozygotic twins share almost 100% of their genetic material while dizygotic twins about 50%. The environment is typically divided into the part of the environment that affects both twins in the same way (the shared environment, C) and the part of the environment that affects one twin but not the other (the unique environment, U) (Silventoinen et al. 2003). Hence, in the absence of interaction and correlation between C and U , we have:

$$E = C + U \quad (6)$$

Assuming epistasis effects to be negligible (a common assumption in twin studies), by inserting Equation (6) into Equation (5), we have:

$$\begin{aligned} Cov(P_{T1}, P_{T2}) &= Cov(A_{T1} + D_{T1} + C_{T1} + U_{T1}, A_{T2} + D_{T2} + C_{T2} + U_{T2}) = \\ &Cov(A_{T1}, A_{T2}) + Cov(D_{T1}, D_{T2}) + Cov(C_{T1}, C_{T2}) + Cov(U_{T1}, U_{T2}) \quad (7) \end{aligned}$$

where indexes “T1” and “T2” represent the two twins for each twin pair studied.

$Cov(P_{T1}, P_{T2})$ is the covariance between the phenotypic values of one twin with the other.

Because each twin's unique environment by definition is independent of that of the other twin, $Cov(U_{T1}, U_{T2})$ is zero for both monozygotic and dizygotic twins. Given that variance is a special case of covariance when the two variables are identical, and that for monozygotic twins A_{T1} , D_{T1} , and C_{T1} equal to A_{T2} , D_{T2} , and C_{T2} respectively, we can formulate the equation from Equation (7) as follows:

$$Cov_{MT}(P_{T1}, P_{T2}) = V_A + V_D + V_C \quad (8)$$

where $Cov_{MT}(P_{T1}, P_{T2})$ is the covariance between the phenotypic values of monozygotic twin pairs studied.

By contrast, dizygotic twins are expected to share half of their genes, which means that the covariance between the phenotypic values of one twin with the other of dizygotic twin pairs studied ($Cov_{DT}(P_{T1}, P_{T2})$) is expected to be equal to half of the additive genetic variance, a quarter of dominant variance², and all of the shared environmental variance (with $Cov(U_{T1}, U_{T2})$ also to be zero). We have:

$$Cov_{DT}(P_{T1}, P_{T2}) = \frac{1}{2}V_A + \frac{1}{4}V_D + V_C \quad (9)$$

It is classically assumed that V_C in Equation (8) and (9) is the same. That is to say, for both monozygotic and dizygotic twin pairs, it is assumed that the shared environment would act in

² For each given gene with two alleles, the possibility that dizygotic twins have the same genotype is one quarter.

the same way if the pair has been reared together.³ V_C can be cancelled by subtracting Equation (9) from Equation (8). The heritability can then be estimated as follows:

$$h_{bTS}^2 = \frac{2\{Cov_{MT}(P_{T1}, P_{T2}) - Cov_{DT}(P_{T1}, P_{T2})\}}{V_P} = \frac{V_A}{V_P} + \frac{\frac{3}{2}V_D}{V_P} \quad (10)$$

We call h_{bTS}^2 broader-sense heritability (the index “b” is for “broader-sense”) from *twin studies*, because the resulting estimate (which is about 0.8 for height) provides an accurate estimate of neither H^2 nor h^2 , although it is closer to H^2 than to h^2 (Falconer and Mackay 1996, 172). That is to say, it corresponds to a definition of heritability that includes *some* elements of broad-sense heritability but not all of it.

Another often used traditional quantitative method to estimate heritability involves a parent-offspring regression. This method also assumes neither gene-environment interaction nor correlation, the covariance between the height of parents (one or the mean of both) and the mean of their offspring (Falconer and Mackay 1996, 164), equals to additive genetic covariance, dominant covariance (the epistasis covariance is assumed to be small and is not included), plus environmental covariance. Hence, Equation (5) can be formulated as follows:

³ This assumption might be problematic because monozygotic twins are often treated more similarly by their parents than are dizygotic twins, and monozygotic twins are more likely to share a placenta than dizygotic twins. The difficulty can be mitigated by using adoption twin studies in which the environments for twins are random on average. But large adoption twins' data are exceedingly difficult to get (Griffiths 2005).

$$\begin{aligned}
Cov(P_P, P_O) &= Cov(A_P + D_P + I_P + E_P, A_O + D_O + I_O + E_O) = \\
&Cov(A_P, A_O) + Cov(D_P, D_O) + Cov(E_P, E_O)
\end{aligned} \tag{11}$$

where indexes “P” and “O” represent the “parents” and the “offspring”.

Two assumptions are then made. The first one is that there is no dominant effects transmitted from the parents to the offspring assuming the parents are unrelated (Doolittle 2012, 178), which means $Cov(D_P, D_O)$ is nil. Another assumption is that there is no correlation between the parents’ environment and the offspring’s environment so that $Cov(E_P, E_O)$ in Equation (11) is also nil. Given that on average, parents share in expectation 50% of genes with their offspring (parents and offspring share half of their genes), it leaves Equation (11) with a result of half of additive genetic variance ($\frac{1}{2}V_A$). Given V_P , h^2 can be estimated straightforwardly.

But the above two assumptions are problematic. First, the assumption of unrelated parents might be violated because of assortative mating in humans resulting in parents to be more genetically similar than two randomly chosen individuals (Guo et al. 2014). Hence, $Cov(D_P, D_O)$ is likely to be non-nil. Second, because the environments experienced by individuals are likely to be more similar within a family line, $Cov(E_P, E_O)$ might not be nil, either. If we take these two factors into consideration, the covariance of the parents and their

offspring is equal to half of additive genetic variance, *plus* a variance term representing effects due to dominance and similarities between environments. This can be written formally as:

$$\text{Cov}(P_p, P_o) = \text{Cov}(A_p, A_o) + \text{Cov}(D_p, D_o) + \text{Cov}(E_p, E_o) = \frac{1}{2}V_A + V_{D\&EC} \quad (12)$$

where $V_{D\&EC}$ represents the variance due to some dominance and environmental correlation effects between the parents and the offspring studied.

The heritability can then be estimated by doubling the parent–offspring covariance in Equation (12) and dividing the total phenotypic variance of the population as follows:

$$h_{bPOR}^2 = \frac{2\text{Cov}(P_p, P_o)}{V_P} = \frac{V_A}{V_P} + \frac{2V_{D\&EC}}{V_P} \quad (13)$$

For similar reasons as with the heritability estimates from twin studies, we call h_{bPOR}^2 broader-sense heritability (with the index “b” also being for “broader-sense”) from *parent-offspring regression*. Indeed, although it is often assumed that h_{bPOR}^2 represent h^2 (Falconer and Mackay 1996, 147), the resulting estimate (also about 0.8 for height) is broader than h^2 as it can include a component led by dominance variance and environmental correlation between parent and offspring.

To conclude this section, heritability estimates in both twin studies and parent-offspring regression include an extra term when compared to h^2 , but they do not correspond to H^2 . For this reason we regroup them under the term h_b^2 for “broader-sense heritability”, such that:

$$h_b^2 = h^2 + h_{other}^2 \quad (14)$$

where h_{other}^2 is the part of heritability contributed by the extra component(s) representing non-additive variance.

3. Heritability in GWAS.

Although any two unrelated individuals share about 99.5% of their DNA sequences, their genomes differ at specific nucleotide locations (Aguilar and Istrail 2013). Given two DNA fragments at the same locus of two individuals, if these fragments differ at a single nucleotide, they represent two variants of a Single Nucleotide Polymorphism (SNP). GWAS focus on SNPs across the whole genome that occur in the population with a probability larger than 1% which are called common SNPs. If one variant of a common SNP, compared to another one, is associated with a significant change on the trait studied, then this SNP is a marker for a DNA region (or a gene) that leads to phenotypic variation. For a polygenic trait like height, if we can detect all the SNPs that associate with it, then all the DNA difference makers that determine height difference can be located.

The development of commercial SNP chips makes it possible to rapidly detect common SNPs of DNA samples from all the participants involved in a study. By using a series of statistical tests, it can be investigated at the population level whether each SNP associates with

that target trait. The choice of the statistical tests depends on the data available as well as the trait studied. For quantitative traits like height, the most common approach is to make an analysis of variance table and assess whether the mean height of a group with one variant at one nucleotide is significantly different from the group with another variant of the same SNP⁴ (Bush and Moore 2012). With all the SNPs associated with height being detected, data from the HapMap project, which provides a list of SNPs that are markers for most of the common DNA variants in human populations (Consortium, International HapMap 3 2010), is used to map the associated SNPs with common DNA variants. These mapped DNA variants, to be distinguished from DNA variants that do not affect the target trait, have been called “causal variants” (Visscher et al. 2012).

Based on the readings of SNP chips as well as further independent tests for SNPs, the effects of the associated SNPs (markers for causal DNA variants) on the trait can be calculated. By estimating the phenotypic variance contributed by these SNPs and the total phenotypic variance of the population, the heritability of causal DNA variants can be estimated as the ratio of the phenotypic variance caused by all the associated SNPs compared to the total phenotypic variance of the population (Weedon et al. 2008). Since it is common for biologists to assume

⁴ For categorical (often binary disease/control) traits, the association test used involves measuring an odds ratio, namely the ratio of the odds of disease for individuals having a specific variant of a SNP, and the odds of disease for individuals who have another variant at the same locus. If the odds ratio of a common SNP is significantly different from 1, then that SNP is considered to be associated with the disease (Bush and Moore 2012).

that genes are only made up of pieces of DNA, it is thought that the variance obtained from all the causal DNA variants represent exactly the additive genetic variance, and the heritability estimated by GWAS should match narrow-sense heritability (h^2) (Yang et al. 2010; Visscher et al. 2006). However, the assumption that additive genetic effects are solely based on DNA sequences is problematic when faced with the evidence of epigenetic inheritance.

As was mentioned in Section 2, traditional quantitative methods for estimating heritability are based on measuring phenotypic values and genetic relations without reaching the molecular level. The genes are not defined physically, but functionally as heritable difference makers (Falconer and Mackay 1996, 123). In other words, they are theoretical units defined by their effects on the phenotype. With the discovery of DNA structure in 1953, it was thought that the originally theoretical genes were found in the physical DNA molecules. Since then, biologists commonly refer to genes as DNA molecules and this assumption is also made by researchers of GWAS. As [author] claim, this step was taken too hastily. If there is physical material, other than DNA pieces, that can affect the phenotype and be transmitted stably across generations, then it should also be thought to play the role that contributes to additive genetic effects.

Many studies have provided evidence for epigenetic inheritance⁵, namely the stable transmission of epigenetic modifications across multiple generations and affect organism's traits

⁵ We use the notion of “epigenetic inheritance” in the broad sense that refers to the inheritance of phenotypic features via causal pathways other than the inheritance of nuclear DNA (Griffiths and Stotz 2013, 112).

(e.g., Youngson and Whitelaw 2008; Dias and Ressler 2014). A classical example of this is the methylation pattern on the promoter of the agouti gene in mice (Morgan et al. 1999). It shows that mice with the same genotype but different methylation levels display a range of colors of their fur, and the patterns of DNA methylation can be inherited through generations causing heritable phenotypic variations. Epigenetic factors such as self-sustaining loops, chromatin modifications and three-dimensional structures in the cell can also be transmitted over multiple generations (Jablonka et al. 2014). Studies on various species suggest that epigenetic inheritance is likely to be ‘ubiquitous’ (Jablonka and Raz 2009).

The increasing evidence of epigenetic inheritance seriously challenges the restriction of the concept of the gene in the evolutionary sense to be materialized only in DNA. Relying on traditional quantitative methods, it is impossible to distinguish whether additive genetic variance is DNA based or based on other material(s). Some transmissible epigenetic factors, which are not DNA based, might *de facto* be included into the additive genetic variance used to estimate h^2 . This extension of heritable units also echoes to the recent suggestion that genetic (assuming genes to be DNA based) and non-genetic heredity should be unified in an inclusive inheritance theory (Danchin 2013; Day and Bonduriansky 2010).

To apply the idea that some epigenetic factors can lead to additive genetic effects, the additive variance of them ($V_{A_{epi}}$) should be added to the additive variance of DNA sequences (V_{ADNA}) to obtain V_A . Assuming there is no interaction between $V_{A_{epi}}$ and V_{ADNA} , we have:

$$V_A = V_{ADNA} + V_{Aepi} \quad (15)$$

Inserting Equation (15) to Equation (4) leads to:

$$h^2 = \frac{V_{ADNA}}{V_P} + \frac{V_{Aepi}}{V_P} \quad (16)$$

Here we term the first term on the right side of Equation (16) “DNA-based narrow-sense heritability” (h_{DNA}^2), and the second term “epigenetic-based narrow-sense heritability” (h_{epi}^2), we thus have:

$$h_{DNA}^2 = h^2 - h_{epi}^2 \quad (17)$$

4. Dissolving the Missing Heritability.

As we mentioned it in Introduction, since the first successful GWAS was published in 2005 (Klein et al. 2005), there have been a lot of proposals for methodological improvements in GWAS (Manolio et al. 2009; Eichler et al. 2010). Studies have been conducted according to those proposals that permit to obtain higher heritability estimates. Examples include increasing the sample sizes which has resulted in more accurate estimates (e.g., Wood et al. 2014), considering all common SNPs simultaneously instead of one by one which has increased the heritability estimates of height from 0.05 to 0.45 (see Yang et al. 2010), and conducting meta-analyses which can lead to more accurate results when compared to single analysis (see Bush

and Moore 2012). Biologists have also suggested to search for SNPs with lower frequencies than 1% in order to account for a wider range of possible causal variants (Schork et al. 2009).

Aside from these partial improvements, our analysis reveals two reasons explaining away the missing heritability problem: a) In traditional quantitative methods, the heritability estimates include extra terms which are not presented in GWAS; b) In GWAS, heritability is estimated solely from causal DNA variants, while in traditional quantitative methods the additive effects contributed by epigenetic difference (h_{epi}^2) are *de facto* included in the estimates.

These two reasons can be shown formally. Using our terminology, missing heritability (MH) equals to the estimates obtained by traditional quantitative methods (h_b^2) minus the estimates obtained by GWAS (h_{DNA}^2), which are 0.8 and 0.45 respectively in the case of height. Thus we have:

$$MH = h_b^2 - h_{DNA}^2 \quad (18)$$

Replacing h_b^2 and h_{DNA}^2 by the right hand side of Equation (14) and (17), we obtain:

$$MH = h_b^2 - h_{DNA}^2 = h^2 + h_{other}^2 - (h^2 - h_{epi}^2) = h_{other}^2 + h_{epi}^2 \quad (19)$$

Which means that the missing heritability results from the part of heritability originating from epigenetic factors stably transmitted across generations, plus the part of heritability originating from non-additives factors.

Our point that part of the missing heritability can be dissolved by considering non-additive effects echoes to the claim that almost all GWAS to date have focused on additive effects might be a reason for the missing heritability (McCarthy and Hirschhorn 2008). Although there is not enough data to confirm that non-additive effects do explain away some part of missing heritability, this claim appears again and again in discussions on the missing heritability problem (see for instance Maher 2008; Frazer et al. 2009; Gibson 2010; Kong 2010; Moore 2010). Yang et al. (2010, 565) disagree with this claim and respond that “[n]on-additive genetic effects do not contribute to the narrow-sense heritability, so explanations based on non-additive effects are not relevant to the problem of missing heritability.”

We agree with Yang et al. (2010) that non-additive effects do not contribute to h^2 . That said, because the heritability estimates obtained from traditional quantitative methods do not strictly correspond to h^2 but include some non-additive elements, non-additive effects cannot be dismissed as irrelevant for the missing heritability problem, though probably they are relevant in a way that both Yang et al. (2010) as well as their opponents did not consider.

5. Conclusion.

We have provided two ways in which the missing heritability problem can be explained away. First, heritability estimates from traditional quantitative methods (h_b^2) are overestimated when

compared to h^2 . The resulting estimates would be smaller if the non-additive elements were eliminated. Second, heritability estimates from GWAS (h_{DNA}^2) are underestimated when compared to h^2 because they do not take into account the additive effects of epigenetic factors behaving like evolutionary genes. The resulting estimates would be larger if epigenetic factors were taken into account. We have voluntarily stayed away from the question of whether heritability should be defined strictly relative to DNA sequences or if it should encompass any factors behaving effectively like an evolutionary gene. Our inclination is that there is no principled reason to exclude non-DNA transmissible factors from heritability measures, but our analysis does not bear on this choice.

References:

- Aguiar, Derek, and Sorin Istrail. 2013. "Haplotype Assembly in Polyploid Genomes and Identical by Descent Shared Tracts." *Bioinformatics* 29 (13): i352–i360.
- Authors. Forthcoming. "The Evolutionary Gene and the Extended Evolutionary Synthesis." *British Journal for Philosophy of Science*.
- Bush, William S., and Jason H. Moore. 2012. "Genome-Wide Association Studies." *PLoS Computational Biology* 8 (12): e1002822.
- Consortium, International HapMap 3. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58.
- Danchin, Étienne. 2013. "Avatars of Information: Towards an Inclusive Evolutionary Synthesis." *Trends in Ecology & Evolution* 28 (6): 351–358.
- Day, Troy, and Russell Bonduriansky. 2011. "A Unified Approach to the Evolutionary Consequences of Genetic and Nongenetic Inheritance." *The American Naturalist* 178 (2): E18–E36.
- Dias, Brian G., and Kerry J. Ressler. 2014. "Parental Olfactory Experience Influences Behavior and Neural Structure in Subsequent Generations." *Nature Neuroscience* 17 (1): 89–96.
- Doolittle, Donald P. 2012. *Population Genetics: Basic Principles*. Vol. 16. Springer Science & Business Media.
- Downes, Stephen M. 2015. "Heritability." In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University.
- Eichler, Evan E., Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. 2010. "Missing Heritability and Strategies for Finding the Underlying Causes of Complex Disease." *Nature Reviews Genetics* 11 (6): 446–450.
- Falconer, Douglas S., and Trudy F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th edition. Longman: Benjamin Cummings.
- Feil, Robert, and Mario F. Fraga. 2012. "Epigenetics and the Environment: Emerging Patterns and Implications." *Nature Reviews Genetics* 13 (2): 97–109.
- Frazer, Kelly A., Sarah S. Murray, Nicholas J. Schork, and Eric J. Topol. 2009. "Human Genetic Variation and Its Contribution to Complex Traits." *Nature Reviews Genetics* 10 (4): 241–251.
- Furrow, Robert E., Freddy B. Christiansen, and Marcus W. Feldman. 2011. "Environment-Sensitive Epigenetics and the Heritability of Complex Diseases." *Genetics* 189 (4): 1377–1387.

- Griffiths, Anthony JF., Susan R. Wessler, Richard C. Lewontin, William M. Gelbart, David T. Suzuki, and Jeffrey H. Miller. 2005. *An Introduction to Genetic Analysis*. 8th edition. New York: W. H. Freeman.
- Griffiths, Paul, and Karola Stotz. 2013. *Genetics and Philosophy: An Introduction*. Cambridge University Press.
- Guo, Guang, Lin Wang, Hexuan Liu, and Thomas Randall. 2014. "Genomic Assortative Mating in Marriages in the United States." *PLoS One* 9 (11): e112322.
- Jablonka, Eva, Marion J Lamb, and Anna Zeligowski. 2014. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Revised edition. MIT Press.
- Jablonka, Eva, and Gal Raz. 2009. "Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution." *The Quarterly Review of Biology* 84 (2): 131–176.
- Johannes, Frank, Vincent Colot, and Ritsert C. Jansen. 2008. "Epigenome Dynamics: A Quantitative Genetics Perspective." *Nature Reviews Genetics* 9 (11): 883–890.
- Klein, Robert J., Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, and Susan T. Mayne. 2005. "Complement Factor H Polymorphism in Age-Related Macular Degeneration." *Science* 308 (5720): 385–389.
- Maher, Brendan. 2008. "Personal genomes: The Case of the Missing Heritability." *Nature News* 456 (7218):18–21.
- Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, and Aravinda Chakravarti. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–753.
- McCarthy, Mark I., and Joel N. Hirschhorn. 2008. "Genome-Wide Association Studies: Potential next Steps on a Genetic Journey." *Human Molecular Genetics* 17 (R2): R156–165.
- Morgan, Hugh D., Heidi GE Sutherland, David IK Martin, and Emma Whitelaw. 1999. "Epigenetic Inheritance at the Agouti Locus in the Mouse." *Nature Genetics* 23 (3): 314–318.
- Schork, Nicholas J., Sarah S. Murray, Kelly A. Frazer, and Eric J. Topol. 2009. "Common vs. Rare Allele Hypotheses for Complex Diseases." *Current Opinion in Genetics & Development* 19 (3): 212–219.
- Silventoinen, Karri, Sampo Sammalisto, Markus Perola, Dorret I. Boomsma, Belinda K. Cornes, Chayna Davis, Leo Dunkel, Marlies De Lange, Jennifer R. Harris, and Jacob VB

- Hjelmborg.2003. "Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries." *Twin Research* 6 (05): 399–408.
- Turkheimer, Eric. 2011. "Still Missing." *Research in Human Development* 8 (3-4): 227–241.
- Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of GWAS Discovery." *The American Journal of Human Genetics* 90 (1): 7–24.
- Visscher, Peter M., Sarah E. Medland, Manuel AR Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, and Nicholas G. Martin. 2006. "Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings." *PLoS Genet* 2 (3): e41.
- Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. "Heritability in the Genomics Era—concepts and Misconceptions." *Nature Reviews Genetics* 9 (4): 255–266.
- Weedon, Michael N., Hana Lango, Cecilia M. Lindgren, Chris Wallace, David M. Evans, Massimo Mangino, Rachel M. Freathy, John RB Perry, Suzanne Stevens, and Alistair S. Hall. 2008. "Genome-Wide Association Analysis Identifies 20 Loci that Influence Adult Height." *Nature Genetics* 40 (5): 575–583.
- Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, and Zoltán Kutalik. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature genetics* 46 (11): 1173–1186.
- Yang, Jian, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, et al. 2010. "Common SNPs Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42 (7): 565–569.
- Youngson, Neil A., and Emma Whitelaw. 2008. "Transgenerational Epigenetic Effects." *Annual Review of Genomics and Human Genetics* 9: 233–257.

Scientific expertise, risk assessment, and majority voting

Thomas Boyer-Kassem*

*Working paper: comments welcome,
but please do not quote without permission.*

February 29, 2016

Abstract

Scientists are often asked to advise political institutions on pressing risk-related questions, like climate change or the authorization of medical drugs. Given that deliberation will often not eliminate all disagreements between scientists, how should their risk assessments be aggregated? I argue that this problem is distinct from two familiar and well-studied problems in the literature: judgment aggregation and probability aggregation. I introduce a novel decision-theoretic model where risk assessments are compared with acceptability thresholds. Majority voting is then defended by means of robustness considerations.

Keywords: scientific expertise, risk, majority voting, robustness, decision theory

*TiLPS, Tilburg University, The Netherlands. Email: t.c.e.boyer-kassem@uvt.nl

1 Introduction

Scientists are often asked by political institutions to give expert advice on pressing questions. For instance, agencies that regulate medicines regularly resort to expert panels, and national scientific academies give advice to the government or to the assemblies. Even after discussing, scientific experts do not always agree on the answer, and when they do, they may disagree on the justification for this answer. How should decisions that involve risk assessments be taken and justified within scientific expert panels? This is the central question studied in this paper. As a matter of fact, many expert panels take decisions using the majority voting rule. This is for instance the case in advisory committees in the European and in the American agencies that grant medicines authorization, respectively the EMA and the FDA.¹ But is it the best decision rule? Is majority voting *on the final decision* the best way to aggregate different experts' opinions, and to track their reasons? This paper is restricted to cases in which the expert panel is asked to take a decision on only one binary question, for instance to answer the question "Is the risk-benefit ratio of some medicine worth it to be authorized for commercial use?". This simple case is already interesting as it corresponds to many real-life cases: some expert panels are constituted on the sole purpose of answering one specific question, or are asked to answer several but logically unrelated questions — e.g. decisions about different medicines.

To study this problem, I introduce a novel decision-theoretic model. The true/false decision is supposed to be taken by comparing a risk assessment a (typically, a probability) to a risk acceptability threshold t , e.g. "true" if and only if $a < t$. For simplicity, a and t are supposed to be in $[0, 1]$, but any quantity might go.² It is assumed that the n experts agree on the threshold value, but differ in their individual risk assessments a_k ($k = 1, \dots, n$) — or conversely, that they agree on the assessment, but disagree on the threshold value. Typically, the question asked to the expert panel is in the form "Is X's risk below t ?". The problem studied in this paper is to determine how the individual a_k 's should be aggregated in comparison with t , so as to give the group's answer to this question (I shall speak equivalently of the group's decision, or of the group's belief on whether the risk is below t). Compared to probability aggregation theory which studies the aggregation of probabilistic opinions, the novelty of this model lies (i) in the introduction of a threshold comparison which projects probabilities into a binary space, and (ii) in the fact that the group has to take a

¹Cf. Hauray and Urfalino (2007), Urfalino and Costa (2015).

²Real quantities can be mapped to the interval $[0, 1]$, for instance with the function $x \rightarrow 1 - 1/(1 + x)$.

stand on one binary question only, and not on a more complex agenda. Compared to judgment aggregation theory which studies the aggregation of an interconnected set of beliefs, the novelty is that individuals do not just have true/false beliefs but probabilistic ones, even if the group is asked to express a true/false belief in the end. The present problem can be considered as a first bridge between these two existing frameworks. The best decision rule for our binary question is likely to depend on the details of characteristics of the question, of the experts, of the available knowledge, and on other details. My methodological approach is not to conduct a detailed study of particular cases, but to look at features which most (interesting) cases share, so as to find general properties of the best decision rule — what is meant by “best” shall be discussed too.

The main claims of this paper are the following. I argue that the framework of probability aggregation cannot help us solve the present problem (Section 2), because the aggregation problems it considers are too general. For the aggregation of scientific risk assessment on a specific question, a theory of its own is needed, and I try to sketch one here. I then argue that robustness considerations clearly legitimate majority voting on the final decision (Section 3). But when justifications for the decisions are sought, majority voting can lead to inconsistencies and the expert panel should aggregate on the reasons separately, before deriving logically its decision (Section 4). Overall, the case for the majority rule is thus a mixed one.

2 Probability Aggregation and Beyond

A standard requirement for a scientific expert panel is that it provides justifications for its decision. In the present model, the decision has to be consistent with the comparison between the risk assessment and the threshold, so a minimal justification is that the panel has a belief on the risk assessment (as all experts have a belief on the risk assessment, it would be weird that the panel claims to refuse the authorization while not being able to say that it believes that the risk assessment is above the threshold). So, our problem includes as a first step the aggregation of the individual risk assessments $\{a_k\}_{1 \leq k \leq n}$ into a single group assessment a — deeper justifications for the group’s decision are contemplated in Section 4. The group’s decision is supposed to be consistent with this assessment, so pragmatically the easiest way to do so may be for the group to first aggregate the individual assessments, and then compare the result to the threshold.

Majority voting on the decision itself is a standard way for expert groups to take decisions, but it does not proceed in that way. Can it be objected that, within our model, it lacks the requirement that the group should be attributed a belief

on the risk assessment? No, for the following reason. The result of the majority vote is “true” if and only if a majority of agents vote “true”, i.e. if and only if a majority of agents have a numerical assessment below the threshold, i.e. if and only if the median of the agents’ assessments is below the threshold. In other words, the majority voting rule on the decision is equivalent to considering that the group’s assessment is the median of the individual assessments. Hence, majority voting is in the race. What are the other challengers? A standard way to aggregate probabilities is to make averages. The linear average is defined as $\sum_k a_k$, and it can be generalized with weights $\omega_k \geq 0$ and $\sum_k \omega_k = 1$, as $\sum_k \omega_k a_k$, to take into account unequal degrees of expertise on the question.³ Other averages are the geometric average or the harmonic average. Our problem is to determine which probability aggregation rule, followed by the threshold, is the best one in our problem. It is easy to see that these various probability aggregation rules can give different binary decisions for the group.⁴

Pooling probability functions has been studied for several years in the theory of probability aggregation (for surveys, cf. Dietrich and List forth., Martini and Sprenger forth., section 3). Can its results be used to select the best aggregation rule in our problem? I shall argue that unfortunately no. The framework of probability aggregation adopts an axiomatic method: it starts by stating several axioms which appear as desirable properties for the pooling function and then studies which function or aggregation rule, if any, satisfies them. The axioms considered in Dietrich and List’s survey can be expressed in our case as:

- **Independence:** the group’s probability a only depends on the individual probabilities a_k .
- **Unanimity preservation:** if all agents’ probabilities a_k are the same, then the group’s probability a is this one too.
- Three **Bayesian axioms:** if some information is learned by all individuals, then the group’s decision changes by conditionalization on that event.

³It is akin to the iterated Lehrer-Wagner model which, starting from respect weights agent have to one another, provides a single probability for the group. However, the iterated Lehrer-Wagner model, and even more its normative interpretation, have been subjected to many criticisms (for a survey, cf. e.g. Martini and Sprenger forth. section 4). As a descriptive model, it is not useful for the present discussion.

⁴Consider for instance the median and the linear average, with three experts with $a_1 = a_2 = 0.04$, $a_3 = 0.10$, and $t = 0.05$. A majority voting on the decision gives a “true” as two experts on three assess the risk to be below the threshold. The linear average (with equal weights) is 0.06, which is higher than t , so this gives a “false”.

The Independence axiom is automatically satisfied here, because our problem contains only one true/false answer, and there is no other probability on which a could depend. The three Bayesian axioms make sense in cases where the expert panel learns new information. In our problem, however, an extensive discussion has already taken place so no agent learns new information anymore, and the expert panel is not making any new inquiry. So the Bayesian axioms are not relevant in our case, and only the Unanimity preservation axiom expresses a desirable property for the aggregation rule.

An essential point to note is that a very large number of aggregation rules satisfy this axiom: the median, linear averaging, geometric averaging, and so on — actually, any convex function of the a_k . This illustrates the fact that a classical uniqueness result from the probability aggregation literature does not hold anymore: the well-known theorem by McConway 1981 and Wagner 1982, which states that linear averaging functions are the *only* independent and unanimity-preserving functions. The reason is that the theorem requires a set of at least three events, whereas our problem only considers two — e.g. the product is risky, with probability a_k , and the product is not risk, with probability $1 - a_k$. Considering a simpler agenda has widened the set of suitable aggregation rules, and no theoretical result from the literature can be used to pick the best one. More generally, the uniqueness and impossibility results from the theory of probability aggregation are useless for our problem. So, how scientific expert panels should aggregate risk assessments is not a simple problem that can be solved straightforwardly with the existing literature, which has focused on general problems with complex agendas, and has thus neglected more specific yet important questions. In the next section, I discuss other desiderata or axioms that we would like to impose on the aggregation rule.

3 Robustness Matters

Scientific risk assessment is supposed to meet some standards of reliability and objectivity, and the aggregation of these assessments should follow alike standards. In this spirit, I now introduce several new requirements for our aggregation rule. The aggregation rule should be sensitive to the right features of our problem, and not to the parasitic ones. It should favor objective features at the detriment of idiosyncrasies or unwanted values (for an analysis of the concept of objectivity, cf. Douglas 2004 — I refer to some of her distinctions below). In other words, the aggregation rule should be robust to some changes that we regard as irrelevant. In this section, I defend three dimensions of robustness that should be taken into account: the risk metrics, the level of detail, and the presence of strategical agents.

Several probability aggregation rules can be considered: linear averaging, geometric averaging, harmonic averaging, among others. As the forthcoming robustness discussion is similar for all the various averagings, I shall simplify it and consider only linear averaging, which shall be contrasted with the median. \mathcal{R}_a denotes the aggregation rule that compares the threshold with the linear average (which thus stands for other averages), and \mathcal{R}_m the aggregation rule that compares the threshold with the median of the individual assessments (which is equivalent to a majority vote on the decision itself).

3.1 Metrics

The formal model I have introduced relies on a quantitative scale — a and t are given numerical values in $[0, 1]$. How is this scale defined in real cases? My talking about probabilities has been only a matter of simplicity given the reduction of the problem to the $[0, 1]$ interval, and typical cases do not bear on well-defined probabilities or explicit scales. For instance, a standard question posed at an FDA advisory committee is “Does the overall risk versus benefit profile for X support marketing in the US ?”⁵. This question supposes that experts identify the risk versus benefit profile, and determine the value of the threshold under which a marketing is warranted. This can be done in a number of ways, and these are essentially value-laden questions⁶ — what is acceptable or not has to do with extra-scientific values, and may also reflect the fact that an expert is risk-averse or risk-seeking. Overall, it makes sense to suppose that both the metrics scale and the threshold depend on the experts. Conversely, as the aggregation procedure is supposed to take place when the experts have extensively discussed, one can make the simplifying assumption that the same facts are known to all, and thus that the risk assessment is the same for all. In that way, our model actually applies in the setting in which a is common to all experts, but each has her own threshold t_k . The fact that the quantitative risk scale is not uniquely defined can be approached from a mathematical viewpoint: any scale can be reparametrized by applying any continuous bijection from $[0, 1]$ to $[0, 1]$, such as $x \mapsto x^2$.

These points make a hard time for the rule \mathcal{R}_a (and other non linear averagings). First, from a practical viewpoint, the dependence of the risk scale metrics on the expert prevents the use of rules which take as inputs the numerical values of the risk assessments or of the threshold. For instance, is it even possible for a chairman to ask her colleagues “Please tell me your overall risk versus benefit acceptability threshold”

⁵Cf. Urfalino and Costa (2015, p.183).

⁶On the role of values in science more generally, and a critic of the value-free ideal, cf. Douglas 2009.

(or assessment), given that each expert may have her own scale? The rule \mathcal{R}_m , as it is equivalent to majority voting, needs not rely on input individual numerical values, and is thus safe from this criticism. Second, even if these practical difficulties could be overcome, some theoretical difficulties remain. Suppose a common scale has been adopted so that all experts can express their t_k . An aggregation rule that depends on the metrics of that common scale can give different outcomes according to the scale employed, as shown in Table 1. This dependence is a problem: which common scale should be chosen? (This is another aggregation problem!) Note that a variant of this problem exists even with a well-defined probability scale. For instance, let A be the event that a certain risk (e.g. carcinogenic substances in food) is responsible for more than 10 cases of cancer in 100,000 people during 1 year. The experts estimate the probability of A , $p(A)$. Consider now A' the event that the risk is responsible for more than 10 cases of cancer in 100,000 people during 10 years. Call $p(A')$ its probability. If the cancer cases are independent along the years, then $p(A') = 1 - (1 - p(A))^{10}$. Because the relation between $p(A)$ and $p(A')$ is not linear, taking the linear average of the experts assessments on A , and transforming it into an assessment on A' , or taking the linear average of the experts assessments on A' , does not give the same result. Which event A or A' is the more “natural” is not clear, and so much more for the right risk group assessment.

This gives good reasons to consider the following requirement: the aggregation rule should be insensitive to the metrics used to describe the problem, i.e. the assessment and the threshold. What should matter is just the relative position of the a and t_k , not their distance which can be due to some idiosyncratic value-laden judgments. This is requiring that the aggregation rule is more objective, under the sense of value-neutral objectivity as characterized by Douglas (2004, p. 460), which does not mean “free from all value influence” (as judging whether a risk benefit ratio is lower enough is bound to involve a value judgment), but takes a position “that is balanced or neutral with respect to a spectrum of values” (here, the balance is reached by taking into account only relative positions). The metrics robustness excludes the rule \mathcal{R}_a which employs a linear average — Table 1 has shown a counter-

	t_1	t_2	t_3	a	Average t	\mathcal{R}_a	\mathcal{R}_m
x scale	.01	.01	.1	.05	.04	False	False
x^2 scale	0.0001	0.0001	.01	0.0025	0.0034	True	False

Table 1: Example in which the rule \mathcal{R}_a gives different answers depending on the scale. The three experts have different thresholds t_k and a common risk assessment a .

example — but not \mathcal{R}_m which relies on the median.⁷

3.2 Level of detail

Another argument for an aggregation rule that does not rely on a specific metrics comes from considerations of the level of detail in which the problem is described. So far, a continuous scale has been assumed, with numerical assessments in $[0, 1]$. Numerical discrete scales could also be used or even qualitative assessments only — it corresponds to decisions under uncertainty and not under risk. Consider for instance the case of the well-known IPCC Assessment Reports, that formulate a synthesis of existing scientific knowledge on climate change issues. The reports use a standardized vocabulary to express uncertainties, with several scales: some are qualitative (e.g. low/medium/high), others are quantitative (and use probabilities).⁸ The historical trend has been to use more quantitative scales and less qualitative scales, but the latter have the advantage of being easily understandable by non-technical audiences, and thus should continue to be used in the future. Some qualitative and quantitative scales are in an explicit correspondence, as illustrated on Table 2. Writing an IPCC report involves synthesizing large amounts of scientific literature, so co-authors of a chapter may have different beliefs on the uncertainties associated with a finding. Whether they express their beliefs on a qualitative or on a quantitative scale, the way their beliefs are aggregated should be smooth and not vary abruptly (some very precise yet qualitative scales are conceivable), all the more than some explicit correspondence exist (Table 2). This is also a question of historically

⁷The comparability of scales is also discussed in Risse’s (2004) political philosophy work, who also takes it as an argument for majority voting.

⁸Cf. e.g. the last report of the Working Group I, Stocker et al (2013, p. 138-142).

Term	Likelihood of the Outcome
Virtually certain	99–100 % probability
Very likely	90–100% probability
Likely	66–100% probability
About as likely as not	33–66% probability
Unlikely	0–33% probability
Very unlikely	0–10% probability
Exceptionally unlikely	0–1% probability

Table 2: Likelihood terms associated with outcomes used in the Fifth Assessment Report of the IPCC (Stocker et al 2013, p. 142).

consistency when switching from qualitative to quantitative scales.⁹ Thus, a sound requirement is that the aggregation rule extends to formulations with discrete and qualitative scales. As the average of non-numerical and qualitative values is not defined, \mathcal{R}_a does not satisfy this requirement. The median is defined on any kind of scale, and \mathcal{R}_m satisfies the requirement. So only \mathcal{R}_m is robust for the level of detail.

3.3 *Bias and strategical votes*

Not all experts are moved by epistemic goals only, and conflicts of interests can arise. For instance, numerous controversies have surrounded the FDA advisory committees along the years (Urfalino and Costa 2015, p. 168-169.) If a better selection of experts may be the solution, the decision rule used in the expert panel can also reduce the impact of bias agents.¹⁰ With \mathcal{R}_a , an expert can strategically express a much lower risk of a medicine to influence the group's average — with a threshold at 10 %, she might express 0.1% instead of just 9%. The aggregation rule should be insensitive to such a strategical vote manipulation, and this is all the more important as the biased agent may have already influenced other agents during the preceding discussion. \mathcal{R}_m is clearly robust in this sense, as an agent has the same influence whether her probability is just below the threshold or close to 0. This is not so for \mathcal{R}_a . This robustness requirement also makes the aggregation rule more objective, in the sense of detached objectivity (Douglas 2004, p. 459): one's personal values (allegiance to a firm) should not prevail on evidence (e.g. that the probability is 9%, as above).

Overall, the three robustness requirements considered here clearly favor \mathcal{R}_m over \mathcal{R}_a . This provides a substantial justification for the traditional democratic rule in expert panels confronted with a binary decision. This result is a real departure from probability aggregation theory, in which linear averaging is justified on solid grounds. Narrowing the agenda and introducing a threshold has changed the solution to the aggregation problem.

⁹One may object that in the IPCC case the co-authors aggregate beliefs without a threshold comparison for a binary decision. Actually, thresholds are implicit: a finding which confidence is too low may not be mentioned. Anyway, the IPCC example can be seen as a mere illustration of the level of detail problem.

¹⁰Biased and extremist agents have been much studied in the literature of opinion dynamics (cf. for instance in Lorenz's 2007 survey), but not so in the literature of opinion aggregation.

4 Reasons

So far, a simplified model of scientific expert panels has been considered, one in which the group is asked to give a binary decision. As argued, the first step in justifying that decision consists for the panel to have a belief on the risk assessment, which is given by the median of the individual assessments in the case of \mathcal{R}_m . However, expert panels are often asked to provide a deeper justification. The question then arises of how the panel should aggregate its members views on this justification. In this section, I propose a novel but simple model for individual numerical assessment justification, in line with my previous threshold model.

Perhaps the most typical interpretation of the risk assessment a is that of a (subjective) probability. Suppose this probability is determined by m independent factors ($m \geq 2$). For instance, the risk associated with a medicine comes from m unrelated secondary effects. Then a is the probability that at least one risk factor triggers:

$$a = 1 - \prod_{j=1}^m (1 - a_j). \quad (1)$$

Each expert k is supposed to have her own assessment of each factor $a_{k,j}$ ($j = 1, \dots, m$). Our problem is then to aggregate the $n \times m$ matrix of probabilities $a_{k,j}$, and to compare that result with the threshold.

As the m factors are independent, a sound requirement is to aggregate the individual assessments on them separately. How should that be done? Adapting the arguments from the previous section, one is lead to the conclusion that the panel should take the *median* of the individual assessments for each factor. However, there is a fundamental limitation to this, due to the previously mentioned theorem by McConway and Wagner's (cf. Section 2). Here is why. Requiring as above that the aggregation proceeds on each factor independently is just requiring the classical independence axiom. Another legitimate requirement is the classical axiom of unanimity preservation: if all experts agree on the risk assessment for one factor, then the panel should take this value as its own. As $m \geq 2$, all the conditions of the theorem by McConway and Wagner are fulfilled¹¹, so its conclusion apply: the only probability aggregation rule on the set of factors and on the overall decision is linear averaging. This reveals that, if groups use the median to determine both the independence factors' values and the overall risk (according to the above results), then it does not give a probability function and inconsistencies can arise. Table 3

¹¹Each of the $m \geq 2$ factors can be triggered or not, so there are at least 4 events, which is higher than the 3 required in the theorem.

gives such an example. In other words, asking the expert panel to take stands on the reasons for its majority decision can lead it to change its decision.

Does it mean that our robustness defense of the median should be discarded? Not necessarily. The theorem by McConway and Wagner assumes that the experts aggregate their views *both* on the independent factors and on the overall risk assessment. But one can have the experts aggregate their views on the independent factors only. The overall risk assessment is then computed according to Equation 1, and the final decision is logically obtained from a comparison between this value and the threshold. In that way, experts do not vote on the final decision directly. This decision rule is a so-called premise-based rule.¹² Then, the linearity result of McConway and Wagner does not apply any more. The robustness considerations from the previous section do apply at the level of independent factors, and they recommend that the group takes the median of the individual assessments.

The present model of factors has assumed that there exists some common numerical scale, so that taking the median of individual assessments makes sense. However, the previous section has in part argued that such a scale may not always exist. In these cases, the present model of independent factors cannot apply. The theory of judgment aggregation offers a general framework for the aggregation of non-numerical reasons or justifications, with true/false beliefs (for reviews, cf. List 2012, Martini and Sprenger forth.). Applying in detail this framework to our problem of scientific justification would require another paper. A general result from this literature, however, is the discursive dilemma: majority voting on a set of true/false beliefs related in a logical way (here: reasons for the decision) may generate inconsistent collective judgments. This echoes our own finding about the median, which corresponds to majority voting in case of a threshold comparison. So whatever

¹²On this strategy more generally, see Cooke (1991), Bovens and Rabinowicz (2006), Hartmann and Sprenger (2012). Another solution to our problem could be the conclusion-based rule, i.e. aggregate only the views on the conclusion, but this is just like the previous section that we are trying to surpass.

Risk aspect	a_1	a_2	$a = 1 - (1 - a_1) \cdot (1 - a_2)$
Agent #1	0.01	0.01	0.0199
Agent #2	0.02	0.01	0.0298
Agent #3	0.01	0.02	0.0298
Median	0.01	0.01	0.0199 or 0.0298 ?

Table 3: A case in which the rule of the median can lead to inconsistencies. With a threshold at e.g. 0.025, the group's decision could be either true or false.

the scale, majority voting on all parts of the question is in great difficulty, and a premise-based solution should be adopted.

5 Conclusion

This paper has investigated the rationale for the majority rule that is often used in scientific expert panels, when dissent persists after discussion, and has looked for the best decision rule in this context. To this end, I have introduced a threshold probability model for individual decisions. Three main points have been shown in the paper: (1) the standard framework of probability aggregation is unable to solve our problem of risk aggregation. (2) robustness considerations clearly favor majority voting on the decision, i.e. comparing the threshold to the median of the individual risk assessments. (The robustness axioms I have advocated, which have been designed from considerations on scientific expert panel, could in return inspire social choice theory). (3) when a justification of the panel's decision is looked for, the median rule (corresponding to majority voting) can lead to inconsistencies. The promising route is to have the group aggregate on the reasons level, not on the final decision one. This should encourage scientific expert panels to divide questions from a logical viewpoints, and to take decisions on sub-problems instead of voting on the final decision directly. Current practices in advisory committees of the FDA and of the EMA could evolve in this respect. However, these claims have only been shown in quite simple and idealized models of decision-making. Future work is needed to investigate other models. These preliminary results have nonetheless cast some serious doubts on the majority voting rule only applied on the final decision.

Note finally the generality of the proposed model, which goes well beyond scientific expertise: the a and t variables can be interpreted as degrees of beliefs or as utility measures, within an epistemology or an economy framework.

References

- Bovens, Luc and Wlodek Rabinowicz. 2006. "Democratic answers to complex questions. An epistemic perspective". *Synthese* 150: 131-153.
- Cooke, Roger M. 1991. *Experts in Uncertainty. Opinion and Subjective Probability in Science*. Oxford University Press.
- Dietrich, Franz and Christian List. Forthcoming. "Probabilistic Opinion Pooling". In *Oxford Handbook of Probability and Philosophy*, Oxford University Press.
- Douglas, Heather E. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138:453-473.
- Douglas, Heather E. 2009. *Science Policy and the Value-Free Ideal*. University of Pittsburgh Press.
- Hartmann, Stephan and Jan Sprenger. 2012. "Judgment aggregation and the problem of tracking the truth." *Synthese* 187:209-221.
- Hauray, Boris and Philippe Urfalino. 2007. "Expertise scientifique et intérêts nationaux. L'évaluation européenne des médicaments 1965-2000". *Annales HSS* 2: 273-298.
- List, Christian. 2012. "The theory of judgment aggregation: an introductory review." *Synthese* 187:179-207.
- Lorenz, Jan. 2007. "Continuous Opinion Dynamics under Bounded Confidence: A Survey." *International Journal of Modern Physics C* 18, 1819.
- Martini, Carlo and Jan Sprenger. Forthcoming. "Opinion aggregation and individual expertise." In *Scientific collaboration and collective knowledge*, ed. by T. Boyer-Kassem, C. Mayo-Wilson and M. Weisberg, Oxford University Press.
- McConway, Kevin J. 1981. "Marginalization and Linear Opinion Pools." *Journal of the American Statistical Association* 76(374): 410-414.
- Risse, Mathias. 2004. "Arguing for Majority Rule". *The Journal of Political Philosophy* 12(1): 41-64.
- Stocker Thomas .F. et al. 2013. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.
- Urfalino, Philippe and Pascaline Costa. 2015. "Secret-Public Voting in FDA Advisory Committees." In *Secrecy and Publicity in Votes and Debates*, ed. Jon Elster, 165-194. Cambridge University Press.
- Wagner, Carl. 1982. "Allocation, Lehrer models, and the consensus of probabilities." *Theory and Decision* 14: 207-220.

Responsiveness and robustness in the David Lewis signalling game

Carl Brusse and Justin Bruner

October 28, 2016

Abstract

We consider modifications to the standard David Lewis signalling game and relax a number of unrealistic implicit assumptions that are often built into the framework. In particular, we explore realistic asymmetries that exist between the sender and receiver roles. We find that endowing receivers with a more realistic set of responses significantly decreases the likelihood of signalling, while allowing for unequal selection pressure often has the opposite effect. We argue that the results of this paper can also help make sense of a well-known evolutionary puzzle regarding the absence of an evolutionary arms race between sender and receiver in conflict of interest signalling games.

1 Signalling games and evolution

Common interest signalling games were introduced by David Lewis (Lewis, 1969) as part of a game theoretic framework which identified communicative conventions as the expected solutions to coordination problems. In recent years, this has informed a growing body of work on the evolution of communication, incorporating signalling games into an evolutionary game theoretic approach to modelling the evolution of communication and cooperation in humans (Skyrms, 2010; Skyrms, 1996).

As the basis for game theoretic modelling of such phenomena, David Lewis signalling games are attractive in their intuitive simplicity and clear outcomes. They are coordination games of common interest between world-observing senders and action-making receivers using costless signals; in contrast to games where interests may differ and where costly signals are typically invoked. In the standard two-player, two-state, two-option David Lewis signalling game (hereafter the ‘2x2x2 game’), the first agent (signaller) observes that the world is in one of two possible states (state1 or state2) and broadcasts one of two possible signals (signal1 or signal2) which are observed by the second agent (receiver) who performs one of two possible actions (act1 or act2). If the acts match the state of the world (i.e. act1 if state1 or act2 if state2) then the players receive a greater payoff than otherwise.

Most importantly, though, the game theoretic results are unequivocal. There exist two Nash equilibria that are, in Lewis’s words, signalling systems where senders condition otherwise arbitrary signalling behaviour on the state of the world, and receivers act on those signals to secure the mutual payoff. The two

systems only differ on which signal gets to be associated with each state of the world¹. Huttegger (2007) and Pawlowitsch (2008) have shown that under certain conditions a signalling system is guaranteed to emerge under the replicator dynamics, a standard model of evolution to be discussed further in section 4.

Of course the degree to which Lewis' approach makes sense is the degree to which we have confidence in the interpretation and application of such a highly idealised model to the more complex target systems. The obvious worry is that by introducing more realistic features into the model one might break or significantly dilute previous findings on the evolution of signalling.

Not surprisingly, then, recent work on Lewis signalling games has investigated the many ways in which such de-idealizations could occur. Some deviations from the standard Lewis signalling game include: more and varied states of the world, the possibility of observational error or signal error, noisy signals, partial deviation in interest between senders and receivers, the reception of more than one signal, and so on. Many such concerns are dealt with favourably in Skyrms (2010), and in work by others. For example Bruner et al. (2014) generalizes beyond the 2x2x2 case and Godfrey-Smith and Martinez (2013) and Godfrey-Smith (2015) mix signalling games of common interest and conflict of interest. One complication of the Lewis signalling game (particularly important for our purposes) is that signalling systems are not guaranteed in the simple 2x2x2 case when the world is biased. In other words, when the probabilities of the world being in state1 or state2 are not equal, a pooling equilibrium in which no communication occurs between sender and receiver is evolutionarily possible.

2 Symmetry breaking

The focus here will be with the idealisation that sender and receiver are equally responsive in strategic settings. Senders and receivers (in the evolutionary treatment of such games) are two populations of highly abstract and constrained agency roles: all that signallers do on observing the state of the world is send a signal, and the receivers must act as though the world is in one or other of the sender-observable states. Of those two roles, it is the restriction on receivers which is the more problematic.

Imagine for example a forager sighting a prey animal at a location inaccessible to her, but close enough to be acquired by an allied conspecific (who cannot observe the animal). In this case, it is easy for the first forager to slip into the signalling role and execute it, whistling or gesturing to her counterpart. To play the receiver role, however, the second forager has to actually re-orient their attention (to some degree) and attempt to engage in appropriate behaviour for the world-state the first has observed (e.g. prey is to the east or to the west, etc.).

The Lewis signalling model by design is constrained such that the receiver's actions are limited to just those acts associated with the sender's observed world-states. It is of course sensible to begin inquiry with as simple of a model as possible and consider a limited range of responses to stimuli. However, our point is that it is more plausible to make these idealizations for signallers than

¹The other two possible outcomes of the game are 'pooling equilibrium', where the receiver plays act1 or act2 unconditionally.

for receivers. Signals are (by stipulation) cheap and easy to send, yet the actions available to the receiver are less plausibly interpreted as intrinsically cheap and free of opportunity cost.

In addition, the informational states drawn on by sender and receiver are also likely to be very different. Any real-life sender's observation of a world state will likely inform their motivations ('we should catch that animal') to dictate a fairly clear course of action ('try to direct the other agent's behaviour'). But all the receiver gets is a whistle, gesture or other signal which (by stipulation) has no pre-established meaning. The experience of observing a strategically relevant state of the world will typically be richer and more detailed than that of observing a strategically relevant artificial signal. All this leads to two concerns. Firstly, asymmetries in the strategic situations are likely to exist between senders and receivers. Receivers are likely to have locally reasonable options available to them other than those relevant to signaller-observed states of the world, and their responsiveness to the strategic situation is therefore less satisfactorily modelled by the strictly symmetric payoff structures of standard signalling games. Call this the structural responsiveness concern.

Secondly, given the likely differences in informational states, goal-directness, workload and opportunity cost implications of sender and receiver roles, we can expect the mechanisms (cognitive and otherwise) which instantiate them to differ as well, quantitatively and qualitatively. This implies that we should not expect their update-responsiveness in any given game to be equal either. Yet the working evolutionary assumption is that senders and receivers update their strategies in an identical manner, modelled using either learning dynamics or replicator dynamics. Call this the evolutionary responsiveness concern.

3 Hedgehog strategies and update asymmetry

The first of these concerns might sound like an argument for abandoning coordination games and moving toward 'conflict of interest' or 'partial conflict of interest' models. However the issue is more specific than this.

The structural responsiveness concern provides parallel motivation to one of Kim Sterelny's (Sterelny, 2012) concerns about Skyrms (2010) use of the Lewis model. Sterelny asks whether the availability of 'third options' on the part of the receiver might undermine the evolution of signalling even when these third options are less valuable than the payoff for successful coordination. As part of a discussion of animal threat responses, he labels this a 'hedgehog' strategy – taking an action which pays off modestly, regardless of the state of the world. To make this concrete, hedgehogs often roll into a ball in response to predators. This is a stark contrast to the more sophisticated behaviour of vervets, who have specific responses to specific threats. Yet the optimal response a vervet takes to one threat – climb a tree when confronted by a leopard – may lead to total disaster when used in response to another threat, such as an eagle. Hedgehogs avoid such outcomes by 'hedging' unconditionally so as to secure a modest payoff. Translated to signalling games, such a gambit may, in many cases, be more attractive than attempting to respond optimally to a signal².

²It is worth noting here that the 'hedgehog' strategy in this Lewis signalling game is in many ways analogous to the risk dominant 'hare' response in stag hunt games. Playing hare instead of stag allows the agent to avoid disaster, but only guarantees the individual a

This compliments the structural responsiveness concern: receivers (especially) might have other options of value which will stand in competition to those assumed in the standard signalling game. Something like these hedgehog strategies are plausible departures from the idealisation and should be expected on the part of the receiver given a realistic demandingness of the role. The question is whether (as Sterelny suspects) including hedgehog strategies might undermine the robustness of evolution toward signalling systems.

Our second concern pertaining to evolutionary responsiveness parallels a well-known evolutionary hypothesis: the so-called Red Queen effect. In competitive relationships such as predator-prey or parasite-host, the Red Queen hypothesis states that species will be constantly adapting and evolving in response to one another just to “stay in the same place” (Van Valen, 1973). This should also be the case in competitive signalling situations – such as predator-prey signalling systems or courtship displays among conspecifics. Signallers and receivers come to not just update their strategies, but to do so at faster or slower rates depending on the nature of the strategic encounter they are entwined in³.

It might seem that in David Lewis signalling games (as with games of common interest in general) the Red Queen effect should have no role to play. However any realistic interpretation of the Lewis signalling game makes it plausible to consider asymmetry in evolutionary responsiveness as likely, if not the norm. First, as argued, the precise cognitive mechanisms and procedures employed by senders and receivers are likely to be different. Different systems will admit to different degrees of plasticity and evolvability – and will have a different set of cross-cutting tasks and utilities that will place their own demands upon them. Quick and easy signalling responses will have different pathways of update and adaptation than the (typically) more complex set of systems which appropriate receiver responses require.

The consideration of multiple use or adaptive reuse also makes the Red Queen hypothesis salient: it is wildly implausible that entirely separate cognitive systems would evolve to deal with competitive signalling situations and coordination-style situations. Cognitive structures which underpin sender or receiver behaviour will likely be subject to evolutionary pressures from competitive as well as cooperative situations, and the responsive nimbleness of sender and receiver strategies is therefore not guaranteed to be the same. We should not assume that the evolution of sender and receiver strategies always proceeds at the same pace.

Finally, there is at least some evidence of a basic asymmetry between sender and receiver roles in the literature on great ape communication. For example, Hobaiter and Byrne (2014) stress the great sophistication and flexibility on the receiver side of Chimpanzee gestural communication, while Seyfarth and Cheney (2003) discuss about how greater inferential sophistication on the receiver side is a feature of many primate communication systems. While these findings do

mediocre payoff. Thus the issues and trade-offs associated with the hedgehog strategy are general concerns not confined to just the Lewis signalling games. Thanks to [name redacted for review] for helping us better see this connection.

³An example of two groups adapting and evolving at different rates can be found in Richard Dawkin’s discussion of his famous Life-Dinner principle (Dawkins and Krebs, 1979). While we expect both predator and prey to adapt to each other, Dawkins claims the prey species will come to evolve at a faster rate than the predator species due to the different selection pressures exerted on both species. Failing to adapt quickly enough for the predator means going hungry for an extra day, while failing to adapt for the prey means death.

not directly support the structural and evolutionary responsiveness concerns, they show that real-life sender and receiver strategies (in our near biological cousins at least) exhibit important differences, suggesting cognitive asymmetries compatible with those concerns.

In summary then, there is reason to consider two structural modifications to the Lewis signalling game as especially salient to the issue of responsiveness: the addition of ‘hedgehog’ strategies for receivers, and differing rates of change in sender and receiver strategies.

4 The model

The evolutionary model we use as a basis for our analysis is the pure-strategy 2x2x2 David Lewis signalling game, with the two-population discrete-time replicator dynamics.

Exact components of the model include two states of the world (L and R), a world-observing signaller with two possible signals (V1 and V2), and a signal-observing receiver with two possible actions (AL and AR). If the receiver’s action matches the state of the world, then both signaller and receiver get a fixed positive success payoff, otherwise their payoff is zero. Signallers and receivers both have four pure strategies available to them (see table 1).

<i>S1</i>	Signal V_1 if L and signal V_2 if R
<i>S2</i>	Signal V_2 if L and signal V_1 if R
<i>S3</i>	Signal V_1 always
<i>S4</i>	Signal V_2 always
<i>S5</i>	Act A_L if V_1 and act A_R if V_2
<i>S6</i>	Act A_R if V_1 and act A_L if V_2
<i>S7</i>	Act A_L always
<i>S8</i>	Act A_R always

Table 1: Signaller and receiver strategies in the standard 2x2x2 common interest signalling game.

For the evolutionary model, the proportions of the different strategies within sender and receiver populations are initially randomly generated. The fitness of each strategy at a time period t is determined by the composition of the opposing population and the payoff associated with each strategy pairing. The proportion of each strategy at play in the next time period $t + 1$ is determined by the standard discrete-time replicator dynamics. For the sender population this is:

$$X_i(t + 1) = X_i(t) \frac{F_i}{F_S}$$

where X_i is the i th sender strategy, F_i is the fitness of that strategy and F_S is the average sender strategy fitness. Likewise, for receivers:

$$Y_j(t + 1) = Y_j(t) \frac{F_j}{F_R}$$

where Y_j is the j th sender strategy, F_j is the fitness of that strategy and F_R is the average receiver strategy fitness. This is repeated until the populations settle

into an evolutionarily stable arrangement. The update process is deterministic and no randomising or mutations are allowed.

5 Modifications and results

We introduce two novel modifications to this model. First, we add a ‘hedgehog’ action A_H for the receiver. Second, we allow the rate of generational change of senders and receivers to vary relative to one other. In addition, the bias of nature is also varied, and we investigate the effects these three departures from the Skyrms/Lewis idealisation have on the evolutionary stability of signalling equilibria.

Turning to our first modification, the receiver now has three possible actions upon observing the signal: A_L , A_R , and A_H . As before a success payoff of 1 is received by both players in the case that the receiver plays A_L while the world is in state L, or the receiver plays A_R while the world is in state R. A payoff of zero is received if A_L or A_R is played otherwise. A payoff of H is received unconditionally if the receiver plays A_H , where the value of H is between 0 and 1. The sender has four familiar pure strategies, whereas the receiver now has five (for simplicity we omit conditional strategies involving A_H).

To adapt the earlier forager story, we can imagine the sender and receiver as an egalitarian hunting party, and the game as a situation where the sender remotely observes the location of a valuable prey animal (left or right) and calls out to the receiver. The receiver is initially unable to observe the prey but can choose to go left or go right (catching the prey if they go in the matching direction), or alternatively to abandon the hunt in order to obtain a less valuable resource they do not need help from the sender to acquire (the hedgehog strategy). Varying the prior probability of the world is equivalent to it being in a situation where it is systematically more likely that the prey is to the left or the right.

In the simple unbiased 2x2x2 signalling game, one of the two signalling equilibria is guaranteed to be reached under the replicator dynamics. In our notation, these equilibria are S1-R1 and S2-R2. Increasing the bias of the world (i.e. making L more probable than R or vice versa) will undermine this, with an increasing proportion of populations instead collapsing to pooling equilibria. This will occur when there are initially few conditional signalling strategies in the sender population. In such situations, receivers do best to simply perform the act that is most appropriate for the more likely state of the world. The incentive for senders to adopt a signalling system then disappears and the community is locked into a pooling equilibrium.

Not surprisingly, we found a similar effect with the hedgehog strategy as values of H, the payoff for A_H , becomes significant. The hedgehog strategy R5 is an additional unilateral response, and is able to draw some initial populations away from the signalling equilibria when H is in excess of 0.5 (i.e., the average payoff for ‘guessing’). This result, for an unbiased world, is illustrated in Figure 1⁴.

⁴Note that the exact range of this effect, including the point at which the effect becomes significant and the y-intercept, are artefacts of the number of world-states and strategies in the model and therefore not general.

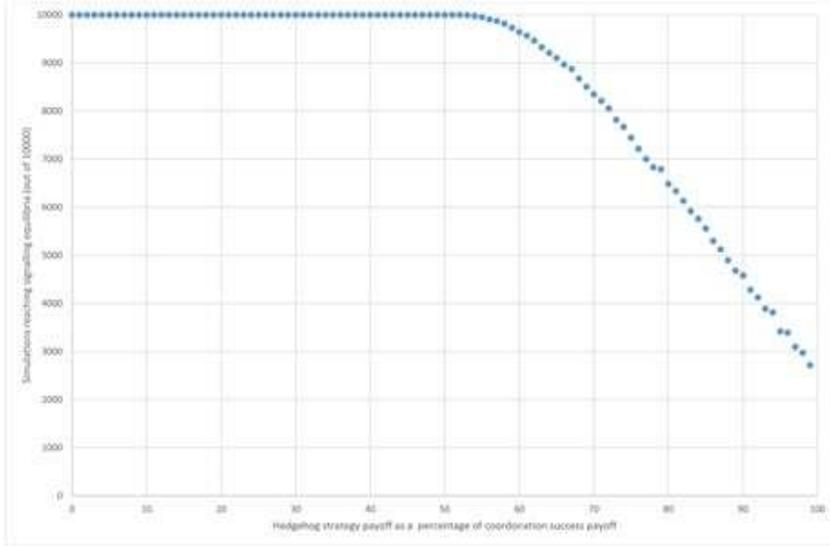


Figure 1: Effect of hedehog payoff on proportion of signalling equilibria.

We observe a more surprising result when the bias and H are varied in combination. Figure 2 shows the results of varying bias for different values of H . The $H = 0$ curve has the expected n-shape, with perfect signalling being degraded as world-bias increases away from the mid-point of even bias between L and R. The inclusion of significant (i.e. $H \geq 0.5$) hedehog payoffs decreases signalling at even bias. As nature becomes increasingly biased, however, the proportion of simulations that head to a signalling system does not go down. In fact we observe a ‘plateau’ followed by a gradual *increase* in the proportion signalling as nature becomes increasingly biased. However, once the bias becomes too extreme, the traditional pooling equilibrium becomes increasingly likely as the payoff associated with simply performing the appropriate act for the more likely state of the world approaches 1. This results in a steep decline in the proportion of simulations that result in signalling systems.

6 Generational asymmetry

We now turn to our second modification of the David Lewis signalling framework in which we introduce a generational asymmetry. We introduced a ‘slow-down factor’ Z to the replicator dynamics in order control the rate at which sender and receiver populations change over time. Composition of the sender and receiver populations are now governed by the following equations:

$$X_i(t+1) = (1 - Z_S)X_i(t) \frac{F_i}{F_S} + X_i(t)Z_S$$

$$Y_j(t+1) = (1 - Z_R)Y_j(t) \frac{F_j}{F_R} + Y_j(t)Z_R$$

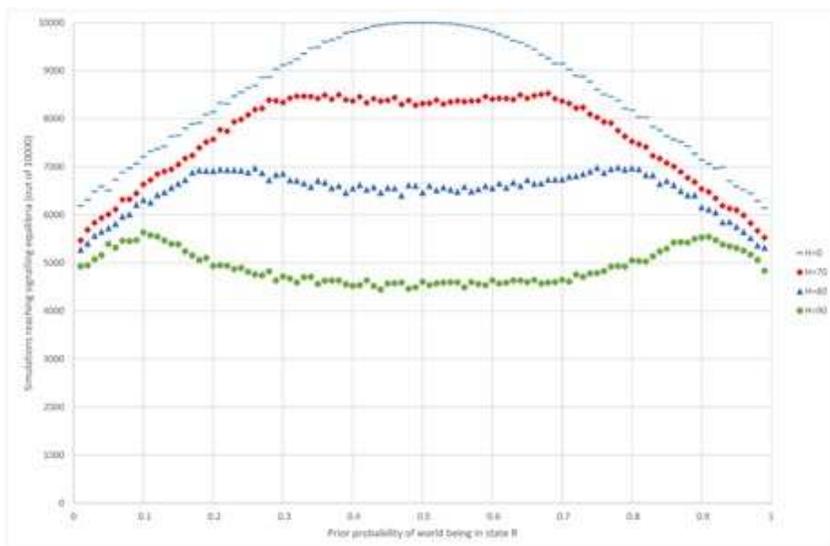


Figure 2: Effect of hedgehog strategy and bias of nature on proportion of signalling equilibria.

Note that when both Z_R and Z_S are zero there is no deviation from the standard replicator dynamics. Rates of changes are slowed as their values increase; for example setting $Z_S = .5$ halves the rate of change for sender strategies. Z_R (alone) being set to 1 means that the composition of the receiver population would not change over time, and only the sender population would evolve.

The result of introducing this generational asymmetry between senders and receivers is that signalling is more likely when sender strategies evolve faster than receiver strategies. This is illustrated in figure 3, where senders (Z_S) and receivers (Z_R) are slowed down to half and one-tenth speeds (with the other population unaltered) as the bias of nature is varied.

Slowing the evolution of the sender population leads to more pooling because, as before, receivers facing a sender population whose conditional signalling is low will begin to gravitate to the act that matches the more likely state of the world (and the threshold for ‘low’ is higher at higher bias). This evolutionary trajectory only reverses if conditional signalling increases rapidly enough to tip the fitness balance toward its matching conditional response, before that response is overpowered. Thus signalling becomes quite a remote possibility when bias is high and senders are slow, occurring in less than 10% of simulations for some parameter values. Slowing the evolutionary responsiveness of the receiver population evolves has the opposite effect – as senders will have time to adopt the best separating strategy given the mix of receiver strategies, and the receiver population slowly adjusts and a robust signalling system establishes. By a similar logic, it is easy to see that a quickly evolving sender population also mitigates against the effect of hedgehog strategies.

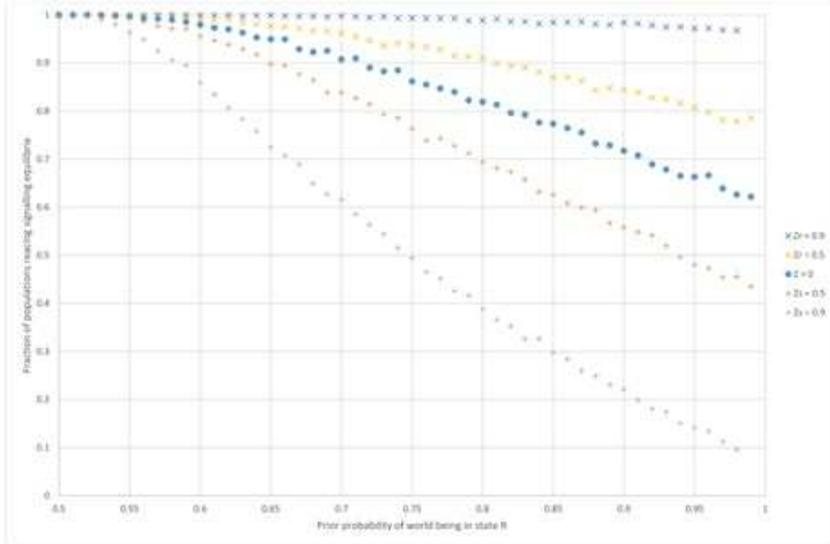


Figure 3: Effect of generational asymmetry and bias of nature on proportion of signalling equilibria.

7 Discussion

We have explored a few well-motivated departures from the highly idealized and simple Lewis signalling game typically considered in the literature. As shown in section 4, breaking the symmetry between senders and receivers often significantly reduces the likelihood that a separating equilibrium emerges. For one, providing receivers with a safe third option which allows them to secure a decent payoff regardless of the state of the world significantly reduces the size of the basin of attraction of the separating equilibrium. Likewise, separating is a remote possibility when receivers outpace senders in the race to adapt.

However the interaction between hedgehog payoffs and bias shows that signalling-undermining effects are not strictly additive. Likewise, the situation is much less bleak when senders evolve at a faster pace than receivers. Interestingly, many scholars in the animal communications literature have noted a similar response asymmetry between sender and receiver in conflict of interest and partial conflict of interest signalling games. For instance, Owren, Rendall, and Ryan (2010) note that senders can easily adapt their signalling behaviour while receivers for the most part have responses to the stimuli produced by senders that are more difficult to change. Thus some have taken to think of signalling as primarily involving the manipulation of receivers by senders.

But this leaves us with an evolutionary puzzle. If there is a conflict of interest between sender and receiver, then what prevents receivers from increasing the speed at which they adapt to the behaviour of the senders? In other words, what explains the absence of an evolutionary arms race between sender and receiver? These are the exact circumstances we would expect the red queen hypothesis to apply. We believe the results of this paper may form the basis of

a novel explanation for this puzzling phenomena. When the interests of sender and receiver are perfectly aligned it is actually in the interest of both parties for the sender population to ‘take the lead’ and evolve at the faster rate, as doing so ensures the community is more likely to hit upon a mutually beneficial signalling system. When the interests of sender and receiver significantly diverge, however, we would expect this not to be the case since both parties now have reason to adapt at a faster pace than the other.

Yet individuals who routinely interact rarely find themselves playing either common interest or conflict of interest signalling games exclusively. As is well known by any parent, not all signalling interactions between relatives are free of conflict. Likewise, agents whose interests are typically thought to be partially opposed, such as two potential mates, may frequently engage in common interest signalling games in contexts unrelated to mating. The point is that a variety of strategic scenarios can hold between sender and receiver, and there is no principled reason to think all interactions will involve perfect alignment or sizable conflict. If so, then a proportion of signalling interactions between sender and receiver may involve no conflict, a partial conflict, or a full conflict of interest. When the proportion of no or low conflict signalling games is significant, the generational asymmetry result from the previous section may hold to some degree. Both sender and receiver will then profit from the sender population evolving at a faster rate than the receiver population, and receivers do best to limit how responsive they are to senders so as to ensure the emergence of informative signalling systems when their interests do overlap. Thus, while it may appear puzzling as to why a receiver is not more responsive when her interests diverge from that of the sender, this confusion might be resolved when the interaction is put into context.

The robustness analysis considered in this paper has in some sense shown how fragile the evolution of signalling can be. Slightly altering the framework in a sensible fashion leads to significantly different results. While many variants of the baseline Lewis signalling game have been explored by philosophers in recent years, more work is required in order to better assess the prospect of signalling in realistic environments.

8 Acknowledgements

We thank Kim Sterelny, Ron Planer and the audiences at the Sydney-ANU Philosophy of Biology Workshop and the 2016 Meeting of the Philosophy of Science Association.

9 Bibliography

Bruner, Justin, Cailin O’Connor, Hannah Rubin, and Simon M. Huttegger. 2014. “David Lewis in the Lab: Experimental Results on the Emergence of Meaning.” *Synthese*, September, 1–19. doi:10.1007/s11229-014-0535-x.

Dawkins, R., and J. R. Krebs. 1979. “Arms Races between and within Species.” *Proceedings of the Royal Society of London B: Biological Sciences* 205 (1161): 489–511. doi:10.1098/rspb.1979.0081.

- Godfrey-Smith, Peter. 1996. *Complexity and the Function of Mind in Nature*. Cambridge; New York: Cambridge University Press.
- Godfrey-Smith, Peter, and Manolo Martínez. 2013. "Communication and Common Interest." *PLoS Comput Biol* 9 (11): e1003282. doi:10.1371/journal.pcbi.1003282.
- Hobaiter, Catherine, and Richard W. Byrne. 2014. "The Meanings of Chimpanzee Gestures." *Current Biology* 24 (14): 1596–1600. doi:10.1016/j.cub.2014.05.066.
- Huttegger, Simon M. 2007. "Evolution and the Explanation of Meaning*." *Philosophy of Science* 74 (1): 1–27.
- Lewis, David K. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.
- Martinez, Manolo, and Peter Godfrey-Smith. 2015. "Common Interest and Signaling Games: A Dynamic Analysis." <http://petergodfreysmith.com/wp-content/uploads/2013/06/Martinez-GS-paper2-Dynamic-Preprint.pdf>.
- Owren, Michael J., Drew Rendall, and Michael J. Ryan. 2010. "Redefining Animal Signaling: Influence versus Information in Communication." *Biology and Philosophy* 25 (5): 755–80. doi:10.1007/s10539-010-9224-4.
- Pawlowitsch, Christina. 2008. "Why Evolution Does Not Always Lead to an Optimal Signaling System." *Games and Economic Behavior* 63 (1): 203–26. doi:10.1016/j.geb.2007.08.009.
- Seyfarth, Robert M., and Dorothy L. Cheney. 2003. "Signalers and Receivers in Animal Communication." *Annual Review of Psychology* 54 (1): 145–73. doi:10.1146/annurev.psych.54.101601.145121.
- Skyrms, Brian. 1996. *Evolution of the Social Contract*. Cambridge University Press.
- . 2010. *Signals: Evolution, Learning, and Information*. Oxford; New York: Oxford University Press.
- Sterelny, Kim. 2012. "A Glass Half-Full: Brian Skyrms's Signals." *Economics and Philosophy* 28 (01): 73–86. doi:10.1017/S0266267112000120.
- Van Valen, Leigh. 1973. "A New Evolutionary Law." *Evolutionary Theory* 1 (1-30). <http://tmtfree.hd.free.fr/albums/files/TMTisFree/Documents/Biology/A>

Experimental Individuation and Retail Arguments

Ruey-Lin Chen

Department of Philosophy, National Chung Cheng University, Taiwan

Jonathon Hricko

Education Center for Humanities and Social Sciences, National Yang-Ming University,
Taiwan

Abstract: Magnus and Callender (2004) argue that we ought to focus on retail arguments, which are arguments regarding the existence of particular kinds of theoretical entities, as opposed to theoretical entities in general. However, *scientists* are the ones who put forward retail arguments, and it's unclear how *philosophers* can engage with such arguments. We argue that philosophers can engage with retail arguments by providing criteria that they must satisfy in order to demonstrate the existence of theoretical entities. We put forward experimental individuation as such a criterion—when scientists experimentally individuate an entity, a realist conclusion about that entity is warranted.

Word Count: 4983

1. Introduction

Magnus and Callender argue that we ought to abandon “wholesale arguments,” which are “arguments about all or most of the entities posited in our best scientific theories” (2004, 321). Instead, we ought to embrace “retail arguments,” which are “arguments about specific kinds of things such as neutrinos, for instance” (2004, 321). This shift in focus rules out standard scientific realism as well as various antirealist positions, and in Section 2, we’ll argue that Magnus and Callender’s position is preferable to these other positions.

However, we recognize that philosophers who choose to abandon wholesale arguments in favor of retail arguments face a potential problem. Dicken (2013) has argued that such philosophers will merely end up repeating the retail arguments that scientists offer. In that case, the turn to retail arguments may entail that no distinctively philosophical work remains to be done. In Section 3, we’ll argue that this is not the case. Not all retail arguments successfully demonstrate the existence of theoretical entities, and it can take some philosophical work to distinguish the ones that do from the ones that don’t.

In Section 4, we’ll put forward a criterion for doing so, which we take from Chen’s (2016) work on experimental individuation. Chen suggests that “[i]f a scientist can realize the individuality of an object in a particular experiment, then she has provided the strongest evidence ... to warrant the reality of the object” (2016, 365). We’ll argue that retail arguments that demonstrate the experimental individuation of a theoretical entity succeed in showing that realism about that entity is warranted.

We'll draw on three examples throughout the paper: Lavoisier's oxygen theory of acidity, J. J. Thomson's work on cathode rays, and Davy's discovery of potassium. We'll conclude, in Section 5, by applying our criterion to these three cases, with the result that the upshot of a retail argument can be either realism, antirealism, or skepticism regarding the existence of a particular kind of theoretical entity.

2. The Turn to Retail Arguments

We'll now introduce Magnus and Callender's position in a bit more detail, and indicate why we take it to be preferable to standard scientific realism (SSR) and antirealism. SSR is a position regarding theories in general—the success of our best theories warrants the claim that they are at least approximately true, as well as the claim that the theoretical entities that they posit exist. Antirealist positions come in a number of different forms, but they all typically endorse claims about theories in general, and deny that success warrants the two claims endorsed by proponents of SSR.

According to Magnus and Callender, there is something that all of these positions have in common, namely, their proponents attempt to support these positions by engaging in wholesale arguments. They focus on two examples of such arguments. First of all, there is the no-miracles argument, according to which the success of our best theories would be a miracle if those theories weren't at least approximately true. Secondly, there is the pessimistic meta-induction, which uses past successful-but-false theories as an inductive basis for concluding that our current successful theories are false as well. The no-miracles argument is taken to support “[w]holesale realism,” which “seeks to explain

the success of science in general”; and the pessimistic meta-induction is taken to support “wholesale anti-realism,” which “seeks to explain the history of science in general” (2004, 321). However, Magnus and Callender argue that these arguments, and wholesale arguments in general, ought to be abandoned. This is because they embody the base rate fallacy, since they don’t take into account the base rate probability of any successful theory being true or false. For this reason, they maintain that wholesale realism and wholesale antirealism ought to be abandoned as well.

Magnus and Callender propose that we ought to replace wholesale arguments with retail arguments. Unlike wholesale arguments, the scope of a retail argument is restricted to a particular theory and/or a particular kind of theoretical entity. By shifting the focus from theories in general to theories in particular, philosophers can *dissolve* the traditional realism debate, with the result that “realism and anti-realism are options to be exercised sometimes here and sometimes there” (2004, 337). This, in turn, opens up the possibility that “[t]here may be good reasons to be a realist about neutrinos, an anti-realist about top quarks, and so on” (2004, 333).

In order to show why this possibility represents an improvement over SSR and antirealism, we’ll now consider a case from the history of chemistry. This case concerns the composition of hydrochloric acid. Scheele was the first to decompose this acid, which he called “acid of salt,” and he identified its constituent substances as phlogiston and “dephlogisticated acid of salt” (1774/1931). However, it was a matter of some controversy whether he had succeeded in decomposing hydrochloric acid. According to Lavoisier’s oxygen theory of acidity, all acids are composed of oxygen (the principle of acidity) and a radical, which can be either a simple substance or a compound (1789/1965,

65, 115). Neither Scheele nor any other chemist had been able to extract the oxygen from hydrochloric acid, which Lavoisier called “muriatic acid.” And so Lavoisier held that it remained undecomposed, and, in accordance with his theory, he hypothesized that it must contain oxygen combined with what he called “the muriatic radical” (1789/1965, 71-72). As for Scheele’s dephlogisticated acid of salt, Lavoisier held that it is a compound of muriatic acid and oxygen, which he called “oxygenated muriatic acid” (1789/1965, 73). Some years later, Davy argued that Scheele was correct, while Lavoisier was in error (1810, 236-37). On Davy’s view, muriatic acid is composed of hydrogen and what he calls “oxymuriatic acid,” which is what Lavoisier called “oxygenated muriatic acid,” and what Scheele called “dephlogisticated acid of salt.” Davy later went on to argue for the elementary nature of this latter substance, and proposed a new name for it: “Chlorine” (1811, 32). His approval of Scheele stems from the fact that Davy, like a number of latter-day phlogiston theorists, identified hydrogen with phlogiston.¹ And the claim that hydrochloric acid is made up of hydrogen and dephlogisticated acid of salt, even if terminologically problematic, is essentially correct. Lavoisier, however, was in error since this acid contains no oxygen, thus falsifying his oxygen theory of acidity.

Proponents of SSR, impressed by narratives of the Chemical Revolution according to which Lavoisier’s oxygen theory defeated the phlogiston theory, are often explicit that their realism applies to the oxygen theory but not to the phlogiston theory.² But in that case, SSR entails the implausible conclusion that Lavoisier’s muriatic radical exists, while Scheele’s dephlogisticated acid of salt does not. It seems much better to

¹ See, e.g., Kirwan (1789, 4-5).

² See, e.g., Hardin and Rosenberg (1982, 610) and Psillos (1999, 291).

conclude that Lavoisier's muriatic radical doesn't exist, while Scheele's dephlogisticated acid of salt does.

Antirealism, at least of the Kuhnian variety, fares no better. Those influenced by Kuhn's (1962/1996) views regarding incommensurability would claim that theoretical entities conceptualized by rival theories should be treated as different entities. However, chemists working in the late eighteenth and early nineteenth centuries shared a set of operations for producing the substance that was variously known as dephlogisticated acid of salt, oxymuriatic acid, and chlorine. It's therefore implausible to maintain that, in light of the fact that these chemists held different theories, they were working with distinct theoretical entities. A trans-theoretical view of the substance that came to be known as chlorine is therefore preferable.

By abandoning wholesale arguments in favor of retail arguments, we can sidestep these difficulties, and simply adopt realism about chlorine (whatever it was called and however it was conceptualized) and antirealism about Lavoisier's muriatic radical. That said, by trading wholesale arguments for retail arguments, we face another difficulty, to which we'll now turn.

3. Can Philosophers Engage with Retail Arguments?

Dicken (2013) has objected that those who abandon wholesale arguments in favor of retail arguments face a serious difficulty. In short, once one does so, it's not clear that any "distinctively philosophical" issues remain to be addressed (2013, 564). Scientists are generally the ones who put forward retail arguments. And if the turn to retail arguments

amounts to merely repeating arguments scientists have offered first, then perhaps nothing distinctively philosophical remains to be done. Our goal in the remainder of the paper is to provide a way of engaging with retail arguments that is distinctively philosophical, and to thereby answer Dicken's objection.

We'll start by considering how scientists demonstrate the existence of theoretical entities, and so we'll now introduce another case from the history of science. This case concerns Thomson's work on cathode rays and his determination of the mass-to-charge ratio (m/e) of the electron. According to the official website of the Nobel Prize, it was because of this work that Thomson "received the Nobel Prize in 1906 for the discovery of the electron, the first elementary particle."³ Thomson (1897, 1906/1967) hypothesized that cathode rays are currents of "carriers of negative electricity" or "corpuscles"—what we now know as electrons.⁴ His hypothesis was not only about the nature of cathode rays, but also about the interaction among cathode rays and other theoretical entities such as electrostatic fields and electrons. In order to determine the mass-to-charge ratio, he measured the deflection of cathode rays passing through an electrostatic field, the strength of the electrostatic field, and other related magnitudes. He interpreted the value that he obtained for m/e in light of his hypothesis, and his experimental results confirmed that hypothesis.

³ Retrieved January 27, 2016 from

<http://www.nobelprize.org/educational/physics/vacuum/experiment-1.html>. See also

Harré (2002) and Whittaker (1989).

⁴ For the identification of Thomson's carriers with electrons, see the reprint of Thomson (1897) in Magie (1969), in which Magie makes the identification.

However, one might ask how it's possible to infer from Thomson's experimental confirmation of his hypothesis to the claim that he had thereby demonstrated the existence of the electron. Philosophers can engage with such a question. And regardless of the answers they provide, they must at least defend those answers by invoking some kind of criterion for concluding that the evidence that scientists have offered does or does not constitute a demonstration of the existence of a given entity. To take one example of such a criterion, Hacking (1983, 23) suggests manipulation: "if you can spray them then they are real." While Thomson manipulated cathode rays, he did not manipulate electrons, and so, according to Hacking's criterion, Thomson did not offer evidence strong enough to demonstrate the existence of electrons.

The important point, for our purposes, is that providing a criterion for granting the reality of a theoretical entity, and determining whether the evidence that scientists have offered satisfies that criterion, constitutes a way for philosophers to engage with retail arguments. Scientists may be the ones who initially put forward retail arguments. But it is a distinctively philosophical task to determine a criterion that can distinguish those retail arguments that demonstrate the existence of a theoretical entity from those that do not. We thus have a way of answering Dicken's objection, provided that, by invoking such a criterion, we are not thereby turning back to wholesale arguments. In the next section, we'll introduce our criterion and argue that applying it does not amount to a wholesale argument.

4. Ontological Commitment and Experimental Individuation

Our proposed criterion for granting the reality of theoretical entities is experimental individuation. A retail argument that demonstrates the experimental individuation of an entity is a good argument for realism about that entity.

Individuation and ontological commitment are connected. When scientists are ontologically committed to the theoretical entities that they posit, this commitment involves not just a belief that the entities exist, but also a responsibility to demonstrate their existence. Demonstrating the existence of a posited entity requires scientists to find an individual instance or sample of that entity, and if a scientist posits a theoretical entity without individuating it, then her ontological commitment is empty.

How do scientists individuate theoretical entities? Answering this question requires us to distinguish *theoretical individuation* from *experimental individuation*. Scientists theoretically individuate an entity if, in the course of theorizing, they describe a set of properties and behaviors of a posited entity by which they can identify it and distinguish it from other entities. However, these descriptions by which scientists theoretically individuate entities require evidence. Scientists can offer evidence for the existence of a theoretical entity if they produce an instance or sample of such an entity by performing an experiment. In doing so, they individuate an entity experimentally.⁵

The relationship between theoretical individuation and experimental individuation is much the same as the relationship between theory and experiment more generally.

⁵ Scientists may also individuate an entity *observationally*, by observing an instance or sample of such an entity. Since observation is itself a complex issue, and since participants in the realism debate rarely question the existence of entities that scientists have observed, we will not discuss observational individuation here.

Various worries about the theory-ladenness of experimentation are relevant here. If a theoretical hypothesis yields a prediction regarding some experimental result, the result may be interpreted in light of the hypothesis. Moreover, since a theoretical hypothesis may involve two or more theoretical entities and their interactions, it can be difficult to show that an experiment produces an instance or sample of the target entity, i.e., that it experimentally individuates that entity. And it can be difficult to judge whether an experiment produces a real individual, as opposed to a mere phenomenon that results from experimental apparatuses and their interactions with experimented objects. For these reasons, a criterion of experimental individuation that is sufficiently independent of theoretical interpretation is needed.

Is there such a criterion for experimental individuation? One candidate is Hacking's manipulation criterion, which we mentioned in Section 3. However, since experimenters can manipulate not just real individuals, but also mere phenomena, manipulation cannot singly serve as the criterion of experimental individuation. Chen (2016) takes Hacking's criterion of manipulation, along with two other criteria, namely, separation and maintenance of structural unity, as jointly constituting a necessary and sufficient condition for the experimental individuation of a theoretical entity. In short, experiments that produce individuals are experiments that separate individuals from their surrounding environment, manipulate them, and maintain their structural unity throughout the process. Importantly, Chen's further conditions ensure that the manipulated object is a real individual as opposed to a mere phenomenon. We take Chen's criteria to offer a satisfactory account of experimental individuation. In Section 5, we'll illustrate his criteria in terms of three retail arguments from the history of science,

and thereby provide some support for our claim that his criteria are satisfactory.

For now, we wish to emphasize two points. First of all, experimental individuation is our proposed criterion for determining whether a retail argument successfully demonstrates the existence of some theoretical entity—it succeeds if it demonstrates the experimental individuation of that entity. Secondly, Chen’s three criteria provide an adequate account of what experimental individuation requires.

Before moving on, we’ll discuss two potential problems with this proposal. First of all, some theoretical entities, like the chemical substances named by mass terms like ‘water,’ ‘phlogiston,’ and ‘oxygen,’ are paradigm cases of non-individuals. It’s therefore not immediately obvious how we can appeal to the notion of experimental individuation when it comes to such entities. We propose to do so by considering the experimental individuation of *samples* of such substances, as we’ll illustrate in Section 5.1, in terms of Davy’s discovery of potassium. Since samples count as individuals, our criterion is applicable to cases involving non-individuals like chemical substances.

Secondly, there’s the issue as to whether the application of our criterion amounts to a kind of wholesale argument. Whether a given retail argument demonstrates the experimental individuation of some theoretical entity is a local matter, grounded in the details of that argument. In contrast, wholesale arguments are not grounded in such local matters. Instead, they rely on claims regarding populations of theories in general, and it is for this reason that they embody the base rate fallacy. We’ve consciously avoided reasoning that may lead to the base rate fallacy. For example, we haven’t argued that the success of our best theories would be a miracle unless the entities they posit can be experimentally individuated. For these reasons, the application of our criterion to retail

arguments does not amount to a kind of wholesale argument. And in that case, we've provided a way of answering Dicken's objection, since our criterion provides a way for philosophers to engage with retail arguments.

5. Application of the Criterion to Three Retail Arguments

Our goal at this point is to show how one can use the criterion we've proposed in order to engage with retail arguments regarding the existence of particular kinds of theoretical entities. We'll discuss three cases: Davy's potassium, Lavoisier's muriatic radical, and Thomson's electron.

5.1 *A Realist Conclusion Regarding Davy's Potassium*

To begin with, we'll argue that Davy demonstrates the experimental individuation of potassium, and thereby provides us with a successful retail argument for realism about that substance.

Davy first isolated potassium by decomposing potash, which he did by means of electrolysis (1808, 4-5). He was the first to decompose potash, though for some time, chemists suspected it to be a compound.⁶ Davy acted on a small piece of moistened potash with a Voltaic battery. As a result, at the negative surface of the battery Davy observed the appearance of "small globules having a high metallic lustre, and being precisely similar in visible characters to quicksilver" (1808, 5). In the lecture in which he

⁶ See, e.g., Lavoisier (1965/1789, 156).

reports these results, Davy goes on to write: “These globules, numerous experiments soon shewed to be the substance I was in search of, and a peculiar inflammable principle the basis of potash” (1808, 5). And later in the lecture, he proposes the name “Potassium [sic]” for the basis of potash (1808, 32).

While this experiment, on its own, does not demonstrate the experimental individuation of a sample of potassium, subsequent experiments that Davy conducted do, and he shows that potassium satisfies all three of Chen’s criteria. First of all, there is Chen’s separation condition: scientists must separate the entities that they produce “from their environments” (2016, 348), and “from the experimental instruments that may have helped produce [them]” (2016, 365). In order to determine whether his results depended on the platinum instruments that he used, Davy performed a number of experiments using a variety of other materials, including copper, silver, and gold (1808, 5). And in order to determine whether his results depended on the fact that he conducted his experiments in the open atmosphere, he performed similar experiments in a vacuum (1808, 5). In all of these cases, he obtained the same results. These experiments collectively show that Davy had separated potassium from its surrounding environment (including the atmosphere and the other components of potash), and from the instruments that he used, thereby satisfying Chen’s separation condition.

Secondly, there is Chen’s condition regarding the maintenance of structural unity. Chen understands structural unity as the idea that “the components of an individual are structured into a whole in some specific manner” (2016, 358). Davy encountered a number of difficulties when it came to maintaining the structural unity of the globules of potassium that he had produced because “they acted more or less upon almost every body

to which they were exposed” (1808, 10). One of the first things Davy notes about the globules is that they did not last long—the ones that did not explode immediately after forming soon lost their metallic luster and became “covered by a white film” (1808, 5). Davy identifies this film as pure potash, and explains how it attracts moisture from the atmosphere, converting the globule into a saturated solution of potash (1808, 7). Eventually, Davy discovered one substance on which potassium did not have much of an effect, namely, recently distilled naphtha (1808, 10). He used that fluid to preserve globules of potassium, and he was able to examine the properties of potassium in the atmosphere by covering the globules with a thin film of naphtha. This method allowed Davy to maintain the structural unity of potassium, thus satisfying Chen’s condition.

Thirdly, there is Chen’s manipulation condition. Chen understands this condition in terms of the “instrumental use” of an object “to investigate other phenomena of nature” (2016, 358). Towards the end of the lecture in which he reports the electrolytic decomposition of potash, Davy conjectures that the globules of potassium he isolated “will undoubtedly prove powerful agents for analysis; and having an affinity for oxygen [sic] stronger than any other known substances, they may possibly supersede the application of electricity to some of the undecomposed bodies” (1808, 44). Making good on this conjecture would amount to showing that chemists can use potassium to decompose previously undecomposed substances, thereby satisfying Chen’s manipulation condition. And in the following year, Davy made good on this conjecture by using potassium to extract the oxygen from a previously undecomposed substance, namely, boracic acid, thereby decomposing it (1809, 76-77).

In sum, Davy shows that samples of potassium satisfy all three of Chen’s criteria.

And by demonstrating the experimental individuation of these samples, Davy presents us with a successful retail argument for realism about potassium.

5.2 An Antirealist Conclusion Regarding Lavoisier's Muriatic Radical

We'll now argue that Davy shows why the experimental individuation of Lavoisier's muriatic radical is not possible, and thereby provides us with a successful retail argument for antirealism about Lavoisier's radical.

As we discussed in Section 2, Lavoisier hypothesized that hydrochloric acid, which he called muriatic acid, is composed of oxygen and a hypothetical substance that he called the muriatic radical. He thereby theoretically individuated the muriatic radical as that substance which combines with oxygen to form muriatic acid, which, in turn, is converted into oxymuriatic acid (i.e., chlorine) by means of combining with even more oxygen. But as we emphasized in Section 4, theoretical individuation is a mere belief, and beliefs require evidence.

Davy (1810, 235-36) provides a retail argument that demonstrates that the experimental individuation of Lavoisier's radical is not possible. He emphasizes the results of various experiments that he and other chemists performed, which show that oxymuriatic acid combines with hydrogen to form muriatic acid. And he goes on to discuss those experiments that seem to show the decomposition of oxymuriatic acid into oxygen and muriatic acid. Davy observes that in these experiments, water is always present. And he concludes that the oxygen that such experiments produce results from the decomposition of the water, not from the decomposition of oxymuriatic acid, which has

not been demonstrated. If oxymuriatic acid doesn't contain oxygen, and muriatic acid contains oxymuriatic acid and hydrogen, then muriatic acid doesn't contain oxygen either. To adopt Davy's later terminology, the only components of muriatic acid are hydrogen and chlorine. Experimentally individuating the muriatic radical would involve separating it from the oxygen with which it combines to form muriatic acid and oxymuriatic acid. And since Davy showed that this is not possible, he gives us a successful retail argument for antirealism about Lavoisier's radical.

5.3 A Skeptical Conclusion Regarding Thomson's Electron

Finally, we'll argue that Thomson neither demonstrates the experimental individuation of the electron, nor shows that it is impossible. Hence, we have an example of an inconclusive retail argument. The proper response to such an argument is skepticism regarding the entity in question, at least until there is a conclusive retail argument regarding the existence of that entity.

Thomson (1897) designed a new type of cathode ray tube (figure 1) to perform a deflection experiment.

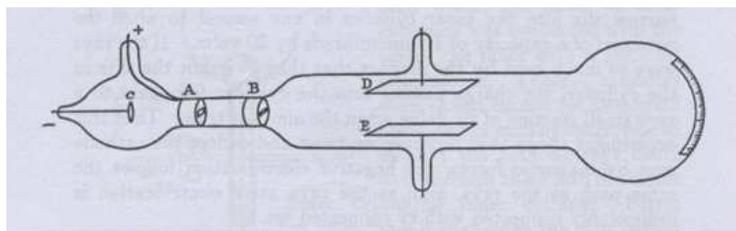


Figure 1. Thomson's cathode ray tube in 1897. Reproduced from Thomson 1969, 586.

This tube contains a cathode *C*, a cylindrical anode *A* with a slit, a cylindrical metal ring *B* with a slit, and a pair of plates *D* and *E* that produce an electrostatic field. A cathode ray is produced when the cathode discharges, and the ray passes through the slits in *A* and *B* before passing through the electrostatic field produced by *D* and *E*. Thomson's goal was to determine whether the ray would be deflected in the field, and to thereby determine the composition of cathode rays. The basic idea was that, if cathode rays were made of ethereal waves, the rays would not be deflected by an electrostatic field; if, however, the rays were made up of negatively electrified bodies, then the rays would be deflected by an electrostatic field.

Thomson's thought was that a cathode would produce both electric currents and cathode rays when discharging, and that, in order to determine the composition of cathode rays, it would be necessary to eliminate the electric currents and experiment with purified cathode rays. Purification is the function of the cylindrical metal ring *B*, which absorbs the electric currents leaked from *A* and thus ensures that the ray passing through *B* is pure. Thomson found that the purified cathode ray was deflected when it passed between the plates *D* and *E*, thus confirming that cathode rays are made up of negatively electrified bodies.

While Thomson satisfies Chen's criteria when it comes to cathode rays, he didn't thereby experimentally individuate the electrons that make them up. Thomson succeeded in *separating* cathode rays from currents; purifying them with the metal ring *B*, and thus *maintaining their structural unity*; and *manipulating* them by deflecting them with an electrostatic field. According to Chen's criteria, one can say that Thomson experimentally individuated cathode rays and demonstrated that they are currents of

negative electricity. But Thomson *presupposed* rather than demonstrated that the currents consist of electrons. He did not demonstrate the existence of electrons, because he did not experimentally individuate them. Hence, the proper response to the retail argument that Thomson gives us is neither realism nor antirealism, but rather skepticism regarding the existence of electrons, at least until there is a conclusive retail argument.

6. Conclusion

Our goal in this paper has been to provide a way for philosophers to engage with retail arguments, and thereby show that, even if we dissolve the traditional realism debate, there is still philosophical work to be done. We've put forward the criterion of experimental individuation in order to determine whether a given retail argument demonstrates the existence of a particular kind of theoretical entity. And we've applied that criterion to three cases, with the result that the upshot of a retail argument can be either realism, antirealism, or skepticism regarding the existence of a particular kind of theoretical entity.

References

Chen, Ruey-Lin (2016). "Experimental Realization of Individuality." In *Individuals Across the Sciences*, ed. Thomas Pradeu and Alexandre Guay, 348-70. New York: Oxford University Press.

Davy, Humphry (1808). "The Bakerian Lecture [for 1807], on Some New Phenomena of Chemical Changes Produced by Electricity, Particularly the Decomposition of the Fixed Alkalies, and the Exhibition of the New Substances Which Constitute Their Bases; And on the General Nature of Alkaline Bodies." *Philosophical Transactions of the Royal Society of London* 98: 1-44.

— (1809). "The Bakerian Lecture [for 1808]: An Account of Some New Analytical Researches on the Nature of Certain Bodies, Particularly the Alkalies, Phosphorus, Sulphur, Carbonaceous Matter, and the Acids Hitherto Undecomposed; With Some General Observations on Chemical Theory." *Philosophical Transactions of the Royal Society of London* 99: 39-104.

— (1810). Researches on the Oxymuriatic Acid, Its Nature and Combinations; And on the Elements of the Muriatic Acid. With Some Experiments on Sulphur and Phosphorus, Made in the Laboratory of the Royal Institution. *Philosophical Transactions of the Royal Society of London* 100: 231-57.

— (1811). The Bakerian Lecture [for 1810]: On Some of the Combinations of Oxymuriatic Gas and Oxygene, and on the Chemical Relations of These Principles, to Inflammable Bodies. *Philosophical Transactions of the Royal Society of London* 101: 1-35.

Dicken, Paul (2013). “Normativity, the Base-Rate Fallacy, and Some Problems for Retail Realism.” *Studies In History and Philosophy of Science* 44(4): 563-70.

Hacking, Ian (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.

Hardin, Clyde L. and Alexander Rosenberg (1982). In Defense of Convergent Realism. *Philosophy of Science* 49(4): 604-15.

Harré, Rom (2002). *Great Scientific Experiments: Twenty Experiments that Changed our View of the World*. New York: Dover.

Kirwan, Richard (1789). *An Essay on Phlogiston and the Constitution of Acids*. 2nd ed. London: J. Johnson.

Kuhn, Thomas S. (1962/1996). *The Structure of Scientific Revolutions*. 3rd ed. Chicago: University of Chicago Press.

Lavoisier, Antoine Laurent (1789/1965). *Elements of Chemistry*. New York: Dover.

Magie, William Francis, ed. (1969). *A Source Book in Physics*. Cambridge, Mass.: Harvard University Press.

Magnus, P. D. and Craig Callender (2004). "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71(3): 320-38.

Psillos, Stathis (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Scheele, Carl Wilhelm (1774/1931). On Manganese or Magnesia; and Its Properties. In *The Collected Papers of Charles Wilhelm Scheele, Translated from the Swedish and German Originals by Leonard Dobbin*, 17-49. London: G. Bell and Sons.

Thomson, Joseph John (1897). "Cathode Rays." *Philosophical Magazine, Fifth Series* 44: 293-316.

— (1906/1967). "Carriers of Negative Electricity. Nobel Lecture, December 11, 1906." In *Nobel Lectures: Physics, 1901-1921*, 145-53. Amsterdam: Elsevier Press.

— (1969). "The Electron." In Magie 1969, 583-97.

Whittaker, Edmund Taylor (1989). *A History of the Theories of Aether and Electricity*.
New York: Dover.

Chirimuuta (forthcoming)

Robustness in Neuroscience

Crash Testing an Engineering Framework in Neuroscience: Does the Idea of Robustness Break Down?¹

ABSTRACT

In this paper I discuss the concept of *robustness* in neuroscience. Various mechanisms for making systems robust have been discussed across biology and neuroscience (e.g. redundancy and fail-safes). Many of these notions originate from engineering. I argue that concepts borrowed from engineering aid neuroscientists in (1) operationalizing robustness; (2) formulating hypotheses about mechanisms for robustness; and (3) quantifying robustness. Furthermore, I argue that the significant disanalogies between brains and engineered artefacts raise important questions about the applicability of the engineering framework. I argue that the use of such concepts should be understood as a kind of simplifying idealization.

“The brain is a physical device that performs specific functions; therefore, its design must obey general principles of engineering.”

Sterling and Laughlin (2015:xv)

1. INTRODUCTION

In this paper I discuss a cluster of issues around the understanding of *robustness* in neuroscience. Systems biologist, Hiroaki Kitano defines

¹ M. Chirimuuta. History & Philosophy of Science, University of Pittsburgh. mac289@pitt.edu. Accepted for presentation at the 2016 Philosophy of Science Association meeting and publication of the proceedings in *Philosophy of Science*.

Chirumuuta (forthcoming)

Robustness in Neuroscience

robustness as, “a property that allows a system to maintain its functions against internal and external perturbations” (Kitano 2004, p.826). According to this definition, in order to determine whether or not a system is robust, one must specify its function, and also specify the kinds of perturbation it faces. Empirically determinable questions then follow about how exactly the system achieves its robustness. Various means for making systems robust have been discussed across biology and neuroscience: copy redundancy, fail-safes, degeneracy, modularity, passive reserve, active compensation, plasticity, decoupling, and feedback (see Figure 1). It is obvious, but still worth emphasising, that most of these notions originate from engineering.

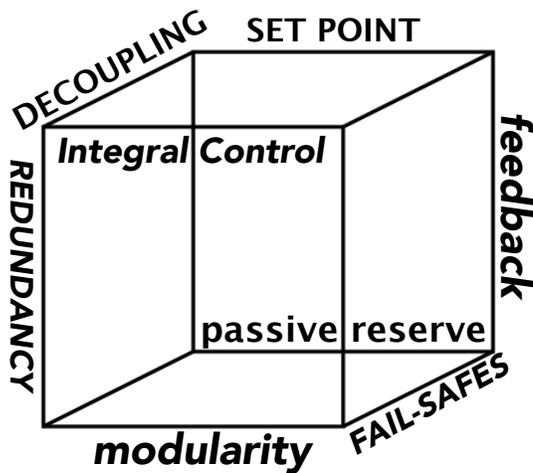


FIGURE 1. The Engineering Framework for Robustness. A set of terms originating from engineering and control theory, which are applied to biological systems to explain how they achieve robust performance.

In Section 2 of this paper I argue that the framework of concepts borrowed from engineering aids neuroscientists in (1) operationalizing robustness by specifying functions of the system and determining possible sources of

Chirimuuta (forthcoming)

Robustness in Neuroscience

perturbation; (2) formulating hypotheses about means for the system to achieve robustness; and (3) showing how robustness may be precisely quantified. This will be shown with examples of neuroscientific research which aims to measure robustness in a retinal circuit (Sterling and Freed 2007), in the motor cortex (Svoboda 2015), and to develop models of homeostatic control (Davis 2006, O’Leary 2014).

In Section 3 I argue that the use of the engineering framework in neuroscience gets stretched, perhaps to breaking point, when applied to systems where (1) there is no principled distinction between processes for robustness and processes which continually maintain the life of the cell; (2) where perturbations are a regular occurrence rather than anomalous events; and (3) where one should not conceive of the system as seeking to maintain a steady state. This point will be illustrated through examination of some recent work from Eve Marder’s laboratory, one of the key centres for research on robustness in neuroscience.

I will argue that the limitations of the engineering notions are put into stark relief when one examines neural systems through the lens of the process approach to biology (Dupré 2012). The engineering perspective, to the extent that it treats biological systems as pre-specified objects with fixed functions, misses many of the features that make robust biological systems fascinating and which are highlighted by the process view.

In Section 4 I will consider if it is necessary to re-engineer the concepts of robustness to be more in line with the dynamicism of biological systems; or alternatively, if we should accept the engineering perspective as it is, as one amongst many idealizing and simplifying heuristics for understanding complex systems like the brain.

2. PUTTING THE ENGINEERING FRAMEWORK TO USE

The robustness of the brain is one of its many extraordinary attributes. By this I mean the fact that brains can undergo moderately severe external perturbations while still maintaining approximately normal function. Obviously, robustness has its limits and the brain's characteristic patterns of resilience and fragility are an important target of research (Sporns 2010, chap. 10). In order to investigate robustness it is necessary first to specify what sorts of perturbations the system is robust to, and then to quantify how robust it actually is. Explanations of robustness can be developed by testing hypotheses concerning the exact mechanisms by which robust performance is achieved. The engineering framework can be put to effective use in each of these processes.

For example, Sterling and Freed (2007) pose the question of how robust the retinal circuit is. They define robustness as the factor by which intrinsic capacity exceeds normal demand, which is the engineer's notion of margin of safety (p.563). The idea can be illustrated through their comparison with bridge design. An engineer designing a road bridge will consider both the anticipated normal demand (e.g. commuter traffic) as well as the unusual demands that might occasionally be placed on the bridge (e.g. the passage of a 30 ton military vehicle). The unusual demand can be thought of as a "perturbation" in Kitano's terms. A robust design will ensure that the system does not break when pushed beyond normal conditions. For a bridge this can be achieved with passive reserve (using thicker steel than is needed under normal conditions) and redundancy (including additional beams so that there are back-up structures if any parts are compromised).

Chirimuuta (forthcoming)

Robustness in Neuroscience

Sterling and Freed take the bridge case to be analogous to the retinal circuit. Normal demand, for the retina, is the intensity of illumination that the eye will encounter under naturalistic stimulation conditions. The safety factor is calculated by experimental determination of the maximum illumination level under which neurons in the retina can maintain their ability to signal to downstream neurons. Sterling and Freed (2007, p.570) report that,

“across successive stages in this neural circuit, safety factors are on the order of 2–10. Thus, they resemble those in other tissues and systems. Their similarity across stages also accords with the principle of symmorphosis—that efficient design matches capacities across stages that are functionally coupled....”

Sterling and Freed’s explanation of robustness depends on the notion of passive reserve. For photoreceptor neurons, this is calculated as the number of vesicles of neurotransmitter available in their synapse for continuous signalling at high-rates without restocking of the vesicles (p.565-6). In arriving at their conclusion about retinal safety margins, they argue that there are at least twice as many vesicles as needed under normal stimulation conditions. In this case we have seen that a design approach borrowed from civil engineering plays a clear and striking role in these neuroscientist’s definition, operationalization and explanation of robustness in the retina.

Another example comes from Davis’s (2006) review of work on homeostatic regulation² in the nervous system. As he writes:

² Note that Davis makes a conceptual distinction between robust properties and properties under homeostatic control: “In general, robustness describes a system with a reproducible output, whereas homeostasis refers to a system with a constant output” (2006, p.308). I will ignore this difference for the purposes of the paper since homeostatic systems conform to Kitano’s general definition robust systems.

Chirimuuta (forthcoming)

Robustness in Neuroscience

“Homeostatic control systems are best understood in engineering theory, where they are routinely implemented in systems such as aircraft flight control. Recently, biological signaling systems have been analyzed with the tools of engineering theory....” (p.314)

Accordingly, homeostatic control systems have a number of “required features”: 1) a set point which defines the target output of the system; 2) feedback; 3) precision in resetting the output back to the set point, following a perturbation; and (normally) 4) sensors which measure the difference between the actual output and the set point (p.309).

Thus control theory offers neuroscientists clear and experimentally testable criteria for determining whether a system undergoes homeostatic regulation, by looking for these required features (e.g. the existence of a set point) in a system. The operating conditions of homeostatic regulation, and the biophysical mechanisms of feedback, sensors, etc., are also open to experimental investigation. Reported examples of properties under homeostatic control are muscle excitation at the neuromuscular junction (p.309) and bursting properties of invertebrate neurons (p.311). More recently, O’Leary et al. (2014, p.818) argue that ion channel expression in their simplified model of invertebrate neurons can be understood as an implementation of *integral control*, a standard control-theoretic architecture.

Figure 2 (if space) schematic for integral control

Chirimuuta (forthcoming)

Robustness in Neuroscience

3. CRASH TESTING THE FRAMEWORK

Before considering the question of whether the engineering framework becomes structurally unsound when applied to some kinds of neural systems, I would like to draw our attention to some of its features. The basic ideas are clearly illustrated in Sterling and Freed's (2007) example of the bridge. When one considers the robustness of an engineered artefact like the bridge, it is presupposed that the system is built up from component parts in such a way as to achieve a specific function. The robustness of the bridge is conceptually distinct from its other designed features or functions, and it can trade off against some of them. For example, the more robust the bridge is to the passage of the occasional heavy vehicle, the more expensive it will be to build (because requiring more steel) (p.563). Moreover, the perturbations against which the system is robust are thought of as atypical events, also conceptually distinct from the normal operations of the system.

There is also the tendency to think of robustness as allowing the system, following a perturbation, to return to its initial stable state. Some experiments specifically involve the operationalization of the robustness of a system as the reversion to a prior state. For example, reporting on an experiment in which mouse premotor cortex in one hemisphere was inhibited using optogenetics during the preparation period for the animal's movement, Svoboda (2015)³ writes, that "[t]his preparatory activity is remarkably robust to large-scale unilateral optogenetic perturbations: detailed dynamics that drive specific future movements are quickly and selectively restored by the network." This notion of robustness as the ability of the system to revert to a

³ To my knowledge, these results have not yet been published in a journal. I have contacted the author to find out if the study is under review or in press.

prior functional states is similar to the idea of *homeostasis* as the ability of a system to stabilize some quantity in spite of external changes.

Figure 3 (if space) After Kitano (2004, Figure 1)

Eve Marder's laboratory has carried out a long term investigation into the ability of neurons to maintain stable electrophysiological properties despite continual turnover of the ion channels embedded in the cell membrane which are responsible for its electrical excitability. This research project is one of the central examples of the study of robustness in neural systems. Marder and her collaborators make ample use of the engineering framework when reviewing other results and reporting their findings. For example, O'Leary et al. (2013, p.E2645) write:

“Both theoretical and experimental studies suggest that maintaining stable intrinsic excitability is accomplished via homeostatic, negative feedback processes that use intracellular Ca^{2+} concentrations as a sensor of activity and then alter[s] the synthesis, insertion, and degradation of membrane conductances to achieve a target activity level.”

What is striking about the characterization of electrophysiological stability in the face of ion channel turnover as a kind of robustness in the face of a perturbation (e.g. p.E2651), is the fact that the turnover is just part of the normal physiology of the cell. There is no functional and stable state of the cell in which this turnover does not occur—a fact which these authors also highlight.⁴ This brings our attention to some strains in the application of the engineering framework to this biological system.

⁴ “neurons in the brains of long-lived animals must maintain reliable function over the animal's lifetime while all of their ion channels and receptors are replaced in the membrane over hours, days, or weeks. Consequently, ongoing turnover of ion channels of various types must occur without compromising

Chirumuuta (forthcoming)

Robustness in Neuroscience

In the basic engineering characterisation of robustness, sketched above, perturbations are different from the normal circumstances in which the system is expected to operate. “Perturbation” carries the everyday connotation of an event which throws the system off balance and is deleterious to its normal functioning. We cannot think of the events of ion channel turnover as perturbations in this sense; they are business as usual for the cell.

Furthermore, it is not in the nature of the system to seek to return to a prior, stable arrangement of its parts. A crucial property of the nervous system is its plasticity: the tendency for its component parts and the connections linking them to be continually sculpted by experience. The homeostatic mechanisms which Marder and colleagues investigate need to be understood as maintaining specific properties (such as a cell’s Ca^{2+} concentration) at a certain point, but not (nor do these researchers claim it) some generalised operation for achieving system-wide internal stability (see §4.4).

In the basic engineering conception of robustness, there is a clear conceptual distinction between the features of a system which allow it to carry out its intended function, and those which make the system robust (even if in reality one individual feature can serve both purposes). In the case of the neuron which has continual ion channel turnover and no definite stable state to return to following these “perturbations”, it is not clear that we can make this distinction. A more natural way to think about this and other biological systems is as ones, unlike engineered artefacts, “designed” to keep changing

the essential excitability properties of the neuron” (O’Leary et al. 2013, p.E2645).

Chirimuuta (forthcoming)

Robustness in Neuroscience

and “designed” to maintain functional stability in the midst of this constant change.⁵

The tensions and strains associated with the application of the basic engineering framework to biological systems can be felt more sharply if we appeal to a process metaphysics of biological “things” (Dupré 2012). According to this view, organisms are not substances but *processes*—items whose existence depends on the taking place of certain changes. This highlights the fact that the life of organisms depends on a continual turnover of its component parts, and that the system as a whole, while living, persists longer than its parts. Yet features and functions of the organism remain relatively stable. For example, memories can endure for decades even though the neurons that form them have undergone material change. This stability must be achieved—somehow. And so processes for robustness are not cleanly distinct from the general maintenance processes which keep the organism alive.

The processual nature of neurons is nicely described by Marder and Goaillard (2012, p.563):

“each neuron is constantly rebuilding itself from its constituent proteins, using all of the molecular and biochemical machinery of the cell.”

(and see F n 4)

⁵ This blurring of the lines between mechanisms for robustness and mechanisms for life is highlighted by Edelman & Gally (2001: 13763) in their discussion of the difference between redundancy and degeneracy in biological systems: “the term redundancy somewhat misleadingly suggests a property selected exclusively during evolution, either for excess capacity or for fail-safe security. We take the contrary position that degeneracy is not a property simply selected by evolution, but rather is a prerequisite for and an inescapable product of the process of natural selection itself.” They also discuss another disanalogy between engineered and biological systems—the applicability of “design” talk.

Chirumuuta (forthcoming)

Robustness in Neuroscience

We can contrast this with the substance metaphysics that we usually assume when thinking about engineered artefacts. A bridge or an aeroplane is what it is because of the parts which comprise it. Its existence does not depend on the occurrence of any process. This is not to deny that an expert in the theory of matter might well argue that the steel of the bridge maintains its integrity because of some fundamental processes. The point is that when characterising the robustness of the bridge or the aeroplane we would not resort to such sophistication. Rather, we think of the bridge as a substance and not a process—a steel structure which, in order to maintain its function in the face of perturbation, must resist rather than effect the swapping around of its component parts.

4. EXAMINING REASONS TO RE-ENGINEER

Now that we have noted these disanalogies between biological organisms and engineered things, we ought to worry that the framework borrowed from engineering is misleading when thinking about robustness in the brain and other biological systems. *Is it time to re-engineer our conceptual tools for thinking about robustness to make them more suitable for characterising living things?* In this section I consider four possible answers to this question.

4.1 *No. The terms in the engineering framework are just words that are used to facilitate communication of the neuroscientific results.*⁶

One potential response to the concerns raised in the previous section is that they stem from a superficial fixation on the vocabulary neuroscientists use when writing about their research. Just because the authors discussed above

⁶ A response along these lines was suggested to me by Timothy O'Leary, in conversation.

*Chirimuuta (forthcoming)**Robustness in Neuroscience*

have employed certain words first introduced by engineers, it does not follow that their understanding of neurophysiology is distorted by comparisons with engineering. For example, I mentioned that the word “perturbation” has a negative connotation which makes it seem inappropriate when describing non-pathological and frequent events like ion channel turnover. It could well be that in the context of this research the term takes on a different meaning—for example, as any event that the system cannot directly control,⁷ such as changes in protein configuration due to thermal noise.

I believe that this response is warranted by what we know of the methodology of some of the investigations discussed above, but not all of them. In the case of Sterling and Freed (2007) I was careful to show that the engineering conceptions directly shaped how these neuroscientists operationalized and quantified robustness, and how they identified mechanisms by which robustness is achieved. There is no indication that they used terms such as “safety factor” to mean something radically different in the context of neuroscience.

A very explicit statement of the aim to apply engineering principles directly to the understanding of the premotor cortex comes from Svoboda (2015):

“preparatory activity is distributed in a redundant manner across weakly coupled modules. These are the same principles used to build robustness into engineered control systems. Our studies therefore provide an example of consilience between neuroscience and engineering.”

Thus the convergence between a neurophysiological and the engineering perspective on the mouse motor planning system is taken to be an important result of this study. This echoes Sterling and Laughlin’s (2015, pp. xiii-xv) proposal that enquiring to see how engineering principles are implemented in

⁷ I thank Timothy O’Leary for this suggestion.

Chirumuuta (forthcoming)

Robustness in Neuroscience

neural systems, and the attempt thereby to reverse-engineer the brain, leads to insights not otherwise available through routine data collection.

4.2 *No. The inadequacies you point out with the engineering framework are based on a caricature of mechanical engineering, not the actual complex discipline.*⁸

My characterisation of the engineering framework assumes that mechanical engineering (the design of bridges, aeroplanes and such like) is paradigmatic of the engineering approach in general. But of course there are many different kinds of engineering, from mechanical to electronic to communications and chemical. It could well be that the mismatch between understanding the robustness of a highly dynamic entity like the brain, and the rather static conception of robust objects that falls out of the basic engineering framework is just an artefact of only focussing narrowly on the kind of engineering that is actually furthest away from neuroscience.

It would take me beyond the scope of this short article (and well beyond my own knowledge of the subject) to sketch out the various possible frameworks associated with each field of engineering specifically, and to see which conception of robustness is most suitable for biology. However, what I will say is that there is evidence in the studies discussed above that neuroscientists themselves do sometimes draw from the mechanically based caricature. This is particularly true of Sterling and Freed (2007). In contrast, when Davis (2006) and O'Leary (2014) make direct appeal to engineering they refer specifically to models in control theory.⁹ This invites questions, still, about whether the paradigm examples of controlled systems (e.g. a car driven on

⁸ This concern was raised by Arnon Levy and Timothy O'Leary.

⁹ See also Zhang and Chase (2015) on the physical control system perspective on brain computer interfaces for motor rehabilitation.

Chirimuuta (forthcoming)

Robustness in Neuroscience

cruise control, a Watt governor, or an aeroplane flown on autopilot) are dynamical enough capture the processual nature of the nervous system.

4.3 *Yes. The brain is so different from an engineered artefact that the framework is misleading and inappropriate.*

In Sections 4.1 and 4.2 I discussed two reasons for thinking that we should not be concerned about any radical disanalogy between robustness in biological and engineered systems. While I agree that these are important points to keep in mind, I do not think that they diffuse the fundamental concern that when neuroscientists borrow engineers' terms in order to study robustness, they risk mischaracterising the brain as more like an engineered artefact than it actually is. Is the appropriate conclusion, then, that a neural circuit is so different from a bridge or an aeroplane that the engineering framework is simply misleading and should be discarded?

The best way to make this strong negative case is to consider some historical examples in which reasoning by analogy with engineered systems seems to have lead neuroscientists and theorists astray. One example comes from von Békésy, a physicist and communications engineer who turned his attention to inhibition in the nervous system. In his book *Sensory Inhibition* he notes that there are feedback loops everywhere in nervous system and he asks how it is that system manages to avoid ending up in a dysfunctional oscillatory state (1967, p.25). It seems that von Békésy is importing his understanding of systems containing feedback from engineering, and in that context oscillations are normally problematic and efforts must be made to dampen them. These days neuroscientists seek to understand how oscillations in the healthy brain (i.e. its characteristic patterns of endogenous activity) are actually responsible for cognitive functions, and how these oscillations differ

Chirimuuta (forthcoming)

Robustness in Neuroscience

from the ones associated with pathologies such as epilepsy and Parkinson's disease.¹⁰

Another example is the comparison of the effects of "noise" in brains and artificial signalling systems..... GET EXAMPLE

This is very different from how neuroscientists understand noise today, which begins with the idea that brains evolved under constraints imposed by noisy "components", which has therefore shaped all aspects of neural computation (Faisal et al. 2008). It would be a mistake to think of the brain processing information in the same way as an electronic computer, but with added redundancy to offset the noisiness of individual processing streams.

The cautionary tales just told give some concrete indications of how imposition of the engineering framework on to neural systems can lead to conclusions which in retrospect appear false and misguided. But it would be too hasty infer from these two examples that current work on robustness in neuroscience is of dubious standing whenever it appeals to the concepts of engineering. A more general argument is the following: *the brain is not like a bridge (or a computer, or an aeroplane on autopilot....); therefore whenever neuroscientists appeal to terms borrowed from the analysis of such systems, they risk saying things that are simply false because they fail to notice relevant disanalogies.* This lays all the sceptical cards on the table. In the last part of the paper I attempt to mitigate these worries.

¹⁰ For a scientific overview see Buzsáki (2006). For discussion of philosophical implications, see Bechtel and Abrahamsen (2013). See also Knuuttila and Loettgers (2013, p.160) on a parallel difference across engineering and cell biology, where oscillations are found to have a functional role.

Chirimuuta (forthcoming)

Robustness in Neuroscience

4.4 *No. Use of the engineering framework should be thought of as a simplifying strategy.*

Neuroscientist Steven Rose (2012:61) writes that:

“one of the most common but misleading terms in the biology student’s lexicon is homeostasis....[the] concept of the stability of the body’s internal environment. But such stability is achieved by dynamic responses; stasis is death, and homeodynamics needs to replace homeostasis as the relevant concept”¹¹

This seems to capture the problem that was first noted in Section 3, that we should not be misled by the engineering framework into thinking of neural systems as seeking to maintain an initial stable state. But we also noted that the neuroscientists employing control-theoretic models of homeostatic mechanisms are not thinking of their systems as seeking stability in this very general way. Instead, they are modelling the stability of a specific variable—in the case of O’Leary et al. (2014), the concentration of Ca^{2+} —and investigating the mechanisms by which it is controlled. To this end, it is reasonable to interpret the system as an integral controller (p.818).¹² Thus it is still useful to talk about homeostasis with respect to Ca^{2+} concentration, even while thinking of the system as a whole, and in reality, as a “homeodynamic” one.

¹¹ Compare Sterling (2012) on the concept of *allostasis* – stability through change with an emphasis on predictive regulation. Day (2005) and O’Leary and Wyllie (2011), in contrast, argue that the concept of homeostasis easily accommodates these dynamic and predictive aspects, and that the term *allostasis* is therefore superfluous. It is an interesting question (but beyond the scope of this paper) whether the narrow or wide definition of *homeostasis* is currently more prevalent amongst biologists and neuroscientists.

¹² Note that O’Leary et al. (2014) study of homeostasis is via a *model* of a neuron. But the model is realistic enough that it is expected to shed light on actual biophysical mechanisms.

Chirumuuta (forthcoming)

Robustness in Neuroscience

I think of neuroscientists whose investigation of robustness in the brain is scaffolded by the engineering framework as providing *idealized mechanistic explanations*. Their explanatory target is, for example, the process by which overall neuronal activity level is controlled via regulation of ion channel gene transcription through a Ca^{2+} sensitive feedback loop. This is standard fodder for mechanistic explanation. At the same time, the framework of engineering—in this case the schematic of the integral controller—serves to direct attention to specific parts and processes in the extremely complex cellular machinery and to interpret them in control theoretic terms (sensors, feedback loops, etc.), while bracketing other aspects not immediately relevant to the explanation of robustness.

Bechtel (2015, p.92) has presented the case that:

“mechanisms are [to be] viewed not as entities in the world, but as posits in mechanistic explanations that provide idealized accounts of what is in the world.”

His example is the idealization (understood as “falsehood”) that scientists introduce by putting boundaries around putative mechanisms which in nature do not exist. In the cases explored in this paper, the idealization comes in through the analogical reasoning of treating a neuronal system *as if* it is an engineered artefact. This, like the positing of boundaries, is a useful way to simplify the explanandum. It enables neuroscientists to bracket some of the known facts about the brain’s messy, Heraclitean nature. But it means, perhaps, that there is a stark difference between the brain viewed *sub specie aeternatis* (what some neuroscientists call the “ground truth” of the brain) and viewed *sub specie mechinae* (in the guise of a machine).

ACKNOWLEDGEMENTS

I am greatly indebted to Timothy O’Leary, Nancy Nersessian and Peter Sterling for their feedback on this work. I would also like to thank the

Chirimuuta (forthcoming)

Robustness in Neuroscience

participants of the Fall 2015 workshop on Robustness in Neuroscience for discussion of the ideas behind this paper, and the audience at the Spring 2016 Re-Engineering Biology conference for their questions and comments on it. Both of these events were hosted by the Philosophy of Science Center at the University of Pittsburgh.

REFERENCES

Bechtel, W. and Abrahamsen, A. (2013). Thinking dynamically about biological mechanisms: Networks of coupled oscillators. *Foundations of Science*, 18:707–723

Bechtel, W. (2015). Can mechanistic explanation be reconciled with scale-free constitution and dynamics? *Studies in History and Philosophy of Science Part C*, 53: 84–93.

von Békésy, G. (1967). *Sensory Inhibition*. Princeton, NJ: Princeton University Press.

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford: Oxford University Press

Davis, G.W. (2006). Homeostatic control of neural activity: from phenomenology to molecular design. *Annu. Rev. Neurosci.* 29, 307–323

Day TA (2005). Defining stress as a prelude to mapping its neurocircuitry: no help from allostasis. *Prog Neuropsychopharmacol Biol Psychiatry* 29, 1195–1200

Dupré, J. (2012) *Processes of Life*. Oxford: Oxford University Press

Chirimuuta (forthcoming)

Robustness in Neuroscience

Edelman GM, Gally JA (2001) Degeneracy and complexity in biological systems. *Proc Natl Acad Sci USA* 98(24):13763–13768.

Faisal, A., L. P. J. Selen and D. M. Wolpert (2008) Noise in the Nervous System. *Nature Reviews Neuroscience*. 9:292-303.

Kitano, H. (2004) Biological robustness. *Nature Reviews Genetics*. 5: 826-837.

Knuuttila, T. and A. Loettgers (2013). Basic science through engineering? Synthetic modeling and the idea of biology-inspired engineering. *Studies in History and Philosophy of Science, Part C* 48, 158–169.

Marder, E. and Goaillard, J.-M. (2012) Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*. 7:563-574

von Neumann, J. (2000). *The Computer and the Brain*. New Haven: Yale University Press.

O’Leary T, Williams AH, Caplan JC, Marder E (2013) Correlations in ion channel expression emerge from homeostatic tuning rules. *PNAS*. 110(28): 809–821

O’Leary T, Williams AH, Franci A, Marder E (2014) Cell types, network homeostasis and pathological compensation from a biologically plausible ion channel expression model. *Neuron* 82(4): E2645–E2654.

O’Leary, T. and D. J. A. Wyllie (2011) Neuronal homeostasis: time for a change? *J Physiol* 589.20:4811–4826

Chirimuuta (forthcoming)

Robustness in Neuroscience

Rose, S. (2012). The need for a critical neuroscience. In *Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience*, S. Choudhury and J. Slaby (eds.) Hoboken, NJ: Wiley-Blackwell.

Sporns, O. (2010). *Networks of the Brain*. Cambridge, MA: MIT Press.

Sterling, P. and M. Freed (2007). How robust is a neural circuit? *Visual Neuroscience*, **24**, 563–571.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior* 106:5–15

Sterling, P. and S. B. Laughlin (2015). *Principles of Neural Design*. Cambridge, MA: MIT Press.

Svoboda, K. (2015). Probing Frontal Cortical Networks during Motor Planning. Abstract, Center for the Neural Basis of Cognition, 10 November 2015. <http://www.braininstitute.pitt.edu/event/probing-frontal-cortical-networks-during-motor-planning>

Zhang, Y. and S. M. Chase (2015). Recasting brain-machine interface design from a physical control system perspective. *J Comput Neurosci* 39:107–118

Eight Myths about Scientific Realism

ABSTRACT: Selective realist projects have made significant improvements over the last two decades. Judging by the literature, however, antirealist quarters seem little impressed with the results. Section I considers the selectivist case and its perceived shortcomings. One shortcoming is that selectivist offerings are nuanced in ways that deprive them of features that—according to many—cannot be absent from any realism “worth having”. Section II (the main part of the paper) considers eight features widely required of realist positions, none of them honored by selectivist projects. Modulo those requirements, even if selectivists managed to clear other shortcomings of their project selectivism would still not be a position worth considering. Next the historical background and present credentials of the requirements in question are examined. All are found to rest on myths and confusions about science and knowledge. If this is correct, realists and antirealists should reject the requirements.

I. Background

The antirealist waves of the 1980s stifled naïve realist projects, but they also gave rise to critical realist reactions, particularly a shift in the way theories are accepted at face value from whole constructs to selected “theory-parts” (existence claims, narratives and structures regarding features beyond the reach of unaided perception). Moves in this “selectivist” direction were variously developed in the 1980s and 1990s, most influentially by Leplin (1984), Worrall (1989b), Kitcher (1993), Leplin (1997), and Psillos (1999). Selectivists see in the history of science a past littered not just with failures but also clear successes, especially after the consolidation of methodologies focused on impressive novel prediction in the early 19th century. The successes selectivists point to involve law-like structures all over physics, functional (as opposed to formally “fundamental”) entities like the particles invoked by the kinetic theory of matter, numerous extinct species hypothesized by Darwin and his circle, structures and processes from microbiology, much in Mendelian genetics, myriads of molecular structures, and most of the subatomic entities deemed well-established since the 1950s, along countless causal networks, histories and functional entities in virtually all theories with warrant in terms of impressive novel predictive success. Selectivists thus respond to skeptical readings of the history of science with optimistic readings, which they argue are better justified than Laudan (1981)’s skeptical appeals. History, Leplin (1984) noted early in the debate on selectivism, is not opposed to realism any more than our experience of ordinary objects is unambiguously veridical.

In selectivist terms, successful scientific theories may provide imperfect representations of unobservable aspects of some of their intended domains, but they do get those aspects right to some significant extent—and *that is what matters to a realist stance*. Realism has to do with

having warranted augmentative inference at levels that reach into unobservables, i.e. beyond the level allowed by its contrast position—constructive empiricism.

Developing selectivism into a mature project has not proved easy. The initial criteria proposed for identifying theory-parts worthy of realist commitment were either too vague or picked up through “retrospective” projection of current. As Kyle Stanford (2006) cautioned, mere retrospective projection of current science reflects limitations of human imagination as easily as it does truth-content and can be variously misleading; also, it can be self-serving, and worse still it severely weakens selectivism by giving up the traditional realist goal of identifying the truthful parts of a theory while the theory is still alive. Realists need to develop compelling criteria for prospective projection, applicable to theories in full flight, and over the last decade selectivists have moved imaginatively to respond to this challenge. One promising contribution is a stronger emphasis on impressive novel predictions as a marker of success and truth content. This trend is multiply developed in works that revisit in detail the cases most used by antirealists as springboard exemplars of gross epistemic failure, as well as studies of other seemingly germane cases from the last 200 years (e.g. Saatsi 2005, Saatsi & Vickers 2011, Votsis 2011, Vickers 2013, @@@). While the debate is far from over, upgraded proposals are on view in the selectivist analyses just cited. At the very least, the initial antirealist arguments from radical underdetermination and so-called “skeptical inductions” have been weakened by selectivist challenges to the antirealist arguments at work. Still, many critics join Stanford in thinking that selectivism lacks a convincing realist criterion for prospective identification of theory-parts. As said, promising selectivist developments seem on view in this regard, but there is something else.

Something seems to be making the selectivist project intellectually unattractive in some quarters, independently of the issue about the criterion for theory-parts. There is, in particular, a perception (not least among many sympathizers of realism) that selectivism advances its case at the cost of diluting its realist import, resulting in *a stance “not worth having.”* By the lights of selective realism, an empirically successful theory T contributes significant truths about unobservables but

- (1) typically, what makes T approximately true is that abstract versions of some of its parts are truthful, making the realist stance applicable to selected fragments of T rather than the integral whole initially intended;
- (2) such truth as T contains need not have universal applicability;

- (3) T need not offer literal truth at its most fundamental level;
- (4) the significance of T's central terms is high in unificationist rather than epistemic terms;
- (5) T adds significantly to our knowledge of unobservables in the intended domain, but there is no reason to expect T to be "right for the most part" at *any* level (what matters is that it yields epistemic gain at theoretical levels).
- (6) T may not instantiate uniformly convergent progress towards any "final description;"
- (7) the intelligibility T confers to its intended domain is generally incomplete.

Each of the above tenets clashes head on with widespread assumptions and expectations regarding a realist stance about theories. The latter, many believe, *should* (1) constitute integral wholes, (2) apply universally, (3) give correct theoretical description, (4) have central terms that refer, (5) be, at least, *right for the most part*. (6) display epistemic progress, and (7) offer substantial intelligibility of the intended domains. Behind these expectations about scientific theories and what theoretical claims amount to is a view on what a *realist position worth having* comprises: to be worth having, a realist position must encompass strong versions of most of the listed assumptions. Antirealists (and not a few realists) routinely take these assumptions for granted. This aspect of the debate needs discussion because, as noted, the assumptions in question are clearly at odds with the selectivist strategy, which—generalizing Worrall (2016) a bit—might be *the only viable realist game in town*.

II. Taxing Assumptions

There is a view, shared by numerous scientists, according to which scientific realism cannot be a position worth having unless it encompasses most of the traits listed at the end of the last section. One problem with those traits is that they provide antirealists with fodder for criticizing positions that embrace them and realists for dismissing positions that lack them. Let us consider the listed items in detail.

(1) **Theories as Integral Wholes.** Selectivism rejects the view that theories and conceptual networks are intellectual constructs made of non-separable parts. The integral wholes vision commits realism to nothing less than complete theories. Motivations for it come from at least two

fronts. One includes linguistic holism and/or the statement view of theories, endorsed in the 1960s and 1970s by thinkers as superficially different as Ernest Nagel and Thomas Kuhn. Another motivation, good for a weaker version of the vision, has been the presumption that some concepts are grounded in “metaphysical necessities,” a position widely held in natural science until the early 1900s. In the 19th century it was thought that breaking of a theory into independently assertible parts had drastic limits. A case in point was the need felt for positing an ether of light, as at the time waves were conceived of within a traditional metaphysics that regarded them as propagating disturbances and thus as ontologically dependent entities that *required* the existence of something being disturbed (@@@). Institutional deference towards similarly presumed conceptual necessities is massively lower now. One major inflection point was the acceptance of Einstein’s Special Relativity, which opened the road to changes in both the conception of light and the requirements of intelligibility in physics.

Nobody thinks now that light is completely as Fresnel or Maxwell imagined, yet—having no conceptual links closed to the possibility of scientific revision—there is little question that Fresnel’s theory got many things right, e.g. what might be termed “Fresnel’s Core”: light is made of microscopic transversal undulations, and these undulations follow the Fresnel laws of reflection and refraction. Abstracted from reference to the wave substratum, this schematic part of the theory spells out a descriptive core that all subsequent theories of light have retained. Once conceptual networks are recognized as relations sustained by revisable inductive conjunctions, scientific “good sense” allows shifts in science towards theory-parts cut out from the rest. There is a historical supplement to this. There has never been much serious allegiance to theory “unbreakability” *in scientific practice*. As scientists developed their ideas, virtually all took a realist stance towards just selected parts of a theory at hand while taking a non-realist stance towards other parts (e.g. Newton’s approach towards Kepler’s cosmology and Galileo’s mechanics; 19th century wave theorists towards particle theories of light, Einstein towards Fresnel’s Core, Einstein towards Newtonian mechanics, molecular geneticists towards Mendelian genetics, and so forth). Being selective about what to take at face value in a theory is exactly what selective realists do, also what we all do in ordinary life. The idea that proper theories are unbreakable integral wholes just rests on myth.

(2) **Universality.** Another widespread assumption is that, for realism, proper scientific theories must hold universally. We find this view expressed in e.g. van Fraassen (1980: 86): from a realist perspective, he claims, “a theory cannot be true unless it can be *extended* consistently, without correction, to all of nature”

This request rests on myth. There is no reason to think that interesting theories can be so extended even at the lowest phenomenal level. Generalizations limited to the observable level typically turn out to be true only over restricted ranges, just as with theoretical generalizations. The standards of acceptability should not be arbitrarily raised against scientific theories. So, past successful theories could not be extended consistently, without correction, to all of nature. However, as selectivists show, those theories made significant cognitive gains at significant levels, where various assortments of the theoretical descriptions they licensed remain both accurate and illuminating. The universality objection, it seems, burdens realism with a suicidal demand.

(3) **Truthful description.** Realists are allegedly claim that what a theory T says about entities, properties, relations and processes should be construed literally; and to take a realist stance towards T is to believe that what it says is literally true. This view comprises three major lines: (3a) literalism, (3b) accuracy realism, and (3c) a methodological supplement.

(3a) Like their biblical counterparts, theory-literalists think one mistake in a narrative is one mistake too many. Phlogiston theory got some of its central claims wrong, as did also Fresnel’s theory, Mendel theory, Bohr’s 1913 theory of the hydrogen atom, and countless other theories, so those theories were all completely wrong.

The antirealist uses of literalism are straightforward. If departures from literal accuracy, however small, make theories count as different, then the chances of a scientist ever picking *the* right theory will be wretchedly small (argument of the bad lots). And the probability of conjecturing the one (and only one) truthful theory will be hopelessly small (problem of the base rate). And, so, at any given time, the chances that the one truthful theory is among the as yet “conceived alternatives” will be overwhelmingly low.

Happily for realists, the expectations in (3a) belong in fairy-tales. Scientific theorizing is rarely strictly literalist. Scientists effectively abandoned literalism early in modern times, as they

began to articulate explanatory idealizations that carried an expectation that nothing in nature exactly realized them. For example, the aim of the kinetic theory of matter developed around 1860 was to causally account for approximate empirical laws that had been gathered in the two previous centuries about the macroscopic behavior of gasses (e.g. $PV = nRT$) and materials (e.g. thermal expansion). Crucially, in the case of gases, the accounts invoked structureless point-particles—the so-called “ideal gas”—that the theorists involved did not believe existed in nature. The ideal gas was *explicitly* an idealization, with a two-fold expectation at work: (i) actual gasses are made of non-ideal corpuscles moving at random and located at relatively large distances from one another “on average”; and (ii) the behavior of those actual corpuscles *instantiated that of the ideal gas to a significant degree* within a certain restricted domain. There was no question that ideal gasses literally construed had to be “real” in order to take the theory realistically. Scientific theories are likewise *generally* false in strictly literal fashion. As with maps, the point of realist interest is the extent to which a theory’s depictions match the *intended* domain. Theoretical representations of empirical domains resemble maps far more than they do assertions (e.g. Giere 2006). Selectivists proceed accordingly: taking a realist stance towards a theory T amounts to claiming only that some of the explanations and descriptions distinct to T are correct by *acceptable standards*.

(3b) In mathematized disciplines literalism easily ups its ante. According to a long lived assumption of quantitative exactitude, there are in nature quantities of which concrete systems have definite values, and in a correct theory the claims it makes correspond to the world with total accuracy. This ideal is found in early modern scientists, notably theorists with strong Platonist leanings such as Kepler.

Dear though these expectations of divine accuracy and depth are, they rest on myth. Such correspondence as mathematized theories have to the world is not conditioned to radical accuracy. As Bertrand Russell noted on behalf of sound epistemology,

“Although this may seem a paradox, all exact science is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man. Every careful measurement in science is

always given with the probable error [...] every observer admits that he is likely wrong, and knows about how much wrong he is likely to be.” (1931: 42)

More recently, in a more comprehensive vein, Paul Teller (2015) complains that “accuracy realism” assumes that the quantities invoked by a theory actually refer. But—he notes—this misunderstands the fabric of theoretical representation, because theories generally formulate *idealizations* that burden quantitative attributions with failure of specificity in picking concrete cases. In the narrowest literal sense, the claim “the meter-standard kept in Paris is 1 meter long” may be true only by *definition*—any attempt to check it with absolute precision against any external objective length would be frustrated by, to begin with, ineliminable thermal and quantum mechanical fluctuations. The point is that one-to-one matching makes no sense as a goal in scientific language, given that so many descriptive words in science are intrinsically vague and/or refer to idealizations. Actual reference to lengths presumes just perspectively *acceptable* (never absolute) accuracy. At the lowest empirical levels also, completely exact assertions are generally neither relevant nor true. This connects with a related point, namely, the *irrelevancy* of these literalist and accuracy assumptions to the actual realism/antirealism debate. Shaped by the discussions started in the 1980s, the dispute is now primarily about whether or not warranted augmentative scientific inferences reach into unobservable domains. Ordinary realism about chairs, cats and mountains fails the ideals of radical literalism and accuracy no less than scientific realism.

(3c) The methodological supplement claims that science would be merely an instrumentalist affair unless theorists aim to produce a complete description of the way things are, with scientists as pursuers of God-like reportage (perfect “mirror reflection”): scientific theories advance towards the truth, all the truth, and nothing but the truth (see e.g. Sankey 2008’s discussion of this). Although this position lost much of its ancient appeal in the 18th century, to this day some top theoreticians continue to wax lyrical expressing it, especially in “editorials”.

“The ‘theory of everything’ is one of the most cherished dreams of science. If it is ever discovered, it will describe the workings of the universe at the most fundamental level and thus encompass our entire understanding of nature. It would also answer such

enduring puzzles as what dark matter is, the reason time flows in only one direction and how gravity works. Small wonder that Stephen Hawking famously said that such a theory would be ‘the ultimate triumph of human reason – for then we should know the mind of God’ ”. (New Scientists, 4 March 2010¹)

This colorful supplement lacks warrant if, as selectivists claim, the realist stance can be consistently and fruitfully applied to selected theory-parts.

The realist badge of honor is not awarded for telling the truth, all the truth, and nothing like the truth about anything—let alone reading the mind of God. It is a distinction for *finite cognitive achievements forged with crooked tools*. See also (6) below.

(4). **Realist Significance of the “Central Tenets” of a Theory.** A related common assumption is this: Even if truthful description may have limits, taking a theory T realistically requires commitment to T’s central tenets (i.e. those about the entities, principles and laws that individuate T). In Laudan’s version, realism about T commits to the view that the T’s central terms *successfully refer*.

There is little question that in numerous scientific theories the central terms fail to refer—on this point we all have a debt of gratitude to Laudan. However, once theories are no longer approached as unbreakable wholes the emphasis on central terms wanes. If anything, the reference that matters is that of theory-parts. Then, on the explanatory side, the scientific focus is on the structures of possibilities of its intended domain D. As such, a theory is not exclusively about the entities and relations invoked at the level of its central terms. Primarily the theory is about D, whose relevant entities and structures include those that may be found at intermediate levels of description—like Fresnel’s Core. A theory may thus be individuated by its central tenets, but the latter do not exhaust the theory’s realist import. The appropriate realist focus is those theoretical claims derivable from the theory and for which there is strong evidence (and so a strong expectation of truthfulness), not whether the terms involved are “central”, “intermediate”, or “peripheral”.

¹ “Knowing the mind of God: Seven theories of everything”, New Scientists, 4 March 2010:
<https://www.newscientist.com/article/dn18612-knowing-the-mind-of-god-seven-theories-of-everything/>

(5). **Being “right for the most part”**. Another related assumption links the realist stance towards a theory with the claim that the theory is right “for the most part”. Michael Devitt, for example, voices this assumption when he defines scientific realism as the doctrine according to which “Most of the essential_unobservables of well-established current scientific theories exist mind-independently and mostly have the properties attributed to them by science” (2005: 769). In his view, theories that are well-established theories *by today’s* methodological standards are right *for the most part*.

This supposition sounds reasonable at first hearing but it too seems suicidal for realism. Virtually all the past theories realists want to be realist about seem to have turned out to be wrong “for the most part”—unless “being right” is granted with postmodern largesse. Newtonian mechanics is “right” for a comparatively tiny regime of speeds and fields. Bohr’s theory of the atom gets impressive aspects right but otherwise is wrong for the most part of the entire quantum domain. Mendel’s theory invites a similar reaction. For all we know, our excellent present physics may be wrong for most of the *total universe*. So, scientifically successful theories seem “wrong for the most part”. But they have great realist import, nonetheless. That import comes from the fact that they get right novel *significant unobservable* aspects of their intended domains. As David Bohm urged long ago, piecemeal caution needs to be exercised in one’s realist commitment to the entities, regularities and processes invoked by well-established current scientific theories (1957, Chapter V). Two lines of reasoning in particular support this prudence (@@@): (1) Qualities, properties of matter, and categories of laws expressed in terms of some finite set of qualities and laws are generally applicable only within limited contexts (in terms of ranges of conditions and degrees of approximation). (2) There is no reason to suppose that new qualities and laws will *always* lead to mere correction refinements that converge in some simple and uniform way. This may occur in some contexts and within some definite range of conditions, but in different contexts and under changed conditions the qualities, properties and laws may be quite novel and lead to dramatic effects relative to what previous theorizing would have led to expect. For example, for bodies moving with speeds negligible compared to the speed of light, the laws of relativity lead to small corrections of the laws of Newtonian mechanics. But they also lead to such qualitatively new results as the “rest energy” of matter. Further laws yet to discover may be vastly more bizarre.

(6) **Progress:** The realist expectation that successful science achieves cumulative truth content about unobservables is frequently nailed to the idea that “modern science is converging on a single picture of the world”. Claims along these lines come in several flavors, in particular (a) linear epistemic progressivism and (b) “metaphysical” realism.

(6a) Convergent progress. Léo Errera expressed the idea in his *Botanique Générale* of 1908: “Truth is on a curve whose asymptote our spirit follows eternally².” This expectation has recurrent mystical roots in science. John Herschel, for example, is cited by Marcel de Serres as saying “All human discoveries seem to be made only for the purpose of confirming more strongly the truths come from on high, and contained in the sacred writings³.”

Convergent progressivism runs against a recurrent realization in modern science. As selectivists recognize, successful theories give knowledge but they usually err at numerous levels of description. Successful theories don’t give us everything there is to know about any intended domain, let alone ‘The World.’ Finite sets of simple laws can provide correct descriptions and predictions when we constrain their context enough, notes Bohm (1957), but we should expect unrestricted theories to be false. Many defenders of scientific objectivity have followed suit, stressing the shift from traditional searches for a comprehensive world-view to explicitly perspectival searches for piece-meal knowledge about domains of current scientific interest, leading to assertions of corresponding partiality.

(6b) In no better shape is the claim that realism is committed to the existence of one true and complete description of the world, whose truth bears one-to-one correspondence to ‘mind-independent reality, so that the purpose of science is to discover that description. Critics persuasively dismiss this brand of realism. But no knowledgeable realist has held such a position in generations. It is a thesis recalled from the grave in the late 1970s and 1980s by Hilary Putnam under the label “metaphysical realism,” a view he presented as an example of a hopelessly jumbled project (e.g. Putnam, 1978: 49, and 1990: Preface).

² *Recueil d'Œuvres de Léo Errera: Botanique Générale* (1908), 193. As translated in John Arthur Thomson, *Introduction to Science* (1911): 57

³ Marcel de Serres, 1845. “On the Physical Facts in the Bible Compared with the Discoveries of the Modern Sciences”. *The Edinburgh New Philosophical Journal* (Vol. 38): 260. [239-271]

(7) **Intelligibility:** Another claim often associated with realism is that science aims to provide truthful explanations that make the phenomena at hand intelligible. This condition comes in (a) radical and (b) moderate strengths. The radical version calls for explanations that leave the intellect content and with no further whys. The weak condition calls for explanations that make the target phenomena *more* but not necessarily fully intelligible.

(7a) Leibniz's rationalist objection to Newton's Theory of Gravitation exemplifies the radical version. He complained that if gravity were thought as a real force, then its effect would be a mysterious action at a distance. Leibniz blamed Newton for introducing "occult" forces into science, and until the end of his life Newton hoped to produce a properly "intelligible" account of gravity involving only action by contact interactions—he did not succeed. Modern scientific theories do not provide radical intelligibility. Once Galileo gave up his initial hope of presenting inertial motion as uniform circular motion, the theory of free fall he accepted left open at least as many whys as it closed. Why or how Galileo's mysterious mathematical structures arise in nature? The same goes for subsequent theorizing. Why or how the regularity given as Newton's law of gravitation arises? Why or how Fresnel's Core arise? Why or how the speed of light is a universal invariant? Contemporary fundamental theories fail radical intelligibility just as clearly.

Realists need not worry about this. Calls for radical intelligibility rest on views of cognition now widely recognized as mythical. Barring mystical insight and such, all actual understanding comes with opaque spots. At every scientific stage scientific warrant (and intelligibility) stops somewhere, albeit usually not at the traditional empiricist boundaries. Realism is compatible with suspending judgment about whether a certain theoretical claim correctly describes a fundamental or derivative aspect of nature. This is exemplified in the stance realists take towards e.g. Fresnel's Core, the invariance of light's speed, and fundamental principles in general.

A theory that saves all the known phenomena but whose reliable parts comprise only structures and explanations at phenomenal levels, provides the lowest level of understanding. This makes for a constructive empiricist take, which escapes skepticism by accepting realism about just the theory's empirical substructures. The point here is that radical theoretical intelligibility is not necessary for taking a realist stance towards a theory. From a selectivist

perspective, the key factor for taking a theory-part realistically is not the “intelligibility” it confers but its indispensability for maintaining the theory’s predictive power in the context of current *background knowledge*. Ptolemaic orbits were denied realist interpretation not primarily because they failed the intelligibility requirement—Ptolemaic constructions went out of their way to honor, of all requirements, *intelligibility* (then guided by the Principle of Uniform Circular Motion for heavenly bodies and the Aristotelian arguments for the fixity of the Earth). Rather, Ptolemaic orbits were refused realist interpretation because the epicycles, deferents and equants they invoked were grossly *underdetermined by extant knowledge* (i.e. available data and cosmological principles). Positive evidence for the orbits specifically proposed was lacking.

None of this is not to question the realist relevance of theories that seek to achieve deep understanding. What is denied is that *scientific* realism must embrace radical intelligibility. Radical intelligibility is a trait realism about observables and every day affairs neither honors nor is expected to honor.

(7b) This brings us to cogent versions of the moderate intelligibility condition. Selectivists take a realist stance only towards theory-parts deemed to be both indispensable for the theory’s success and free of compelling specific doubts against them (@@@). That is, the realist stance goes *only* to tenets for which there is strong positive evidence by modern scientific standards. In all the cases highlighted by realists, the selections supported by the strongest level of evidence available make the target domain intelligible well beyond the observable levels. When, by contrast, the positive evidence for a theory does not reach the unobservable explanatory posits that make the relevant phenomena intelligible, then the best stance to take about the theory is not realism but *constructive empiricism*. This clarifies what introductory characterizations of scientific realism get right about the intelligibility condition: A good theory must not have just significant predictive power but must also make the relevant phenomena *intelligible* (Richard DeWitt 2010: 72). If the theory parts that do this lack evidential warrant, then the reasonable stance towards them is constructive empiricism.

(8) **Realism Worth having.** Topping the above assumptions, there is a popular notion to the effect that a realist stance failing to adhere to most of the above requirements is “*not a realism worth having*”. Against this idea, I have argued that none of the listed assumptions is worth

having. Every one of them lacks convincing warrant. Moreover, even if the assumptions did get proper warrant they face a deeper problem: the assumptions are *irrelevant* to the current realism/antirealism debate—they do not expose relevant contrasts between inferences limited to the phenomenal level and inferences that reach into theoretical levels.

In modern science, virtually all interesting augmentative inferences violate the listed assumptions. So, the latter simply and arbitrarily raise the epistemological standards of acceptability against theoretical assertions. If the above considerations are correct, then, realists and antirealists should reject the assumptions examined in this paper—they all rest on counterproductive myths and confusions.

References

- Bohm, David (1957). *Causality and Chance in Modern Physics*. London: Routledge & Kegan Paul Ltd.
- Devitt, Michael (2005). “Scientific Realism”. In *The Oxford Handbook of Contemporary Philosophy*, Frank Jackson and Michael Smith, eds. Oxford: Oxford University Press: 767-91.
- Giere, Ronald N. (2006). *Scientific Perspectivism*. Chicago: University of Chicago Press.
- DeWitt, Richard (2010): *Worldviews*. Malden, MA: Wiley-Blackwell.
- Kitcher, Philip, 1993. *The Advancement of Science*. Oxford: Oxford University Press.
- Laudan, Larry. 1981. “A Confutation of Convergent Realism”. *Philosophy of Science* 48: 19-49.
- _____. 1984. *Science and Values*. Berkeley: University of California Press.
- _____. 1996, *Beyond Positivism and Relativism: Theory, Method and Evidence*. Boulder, CO: Westview Press.
- Leplin, Jarrett, ed. 1984. *Scientific Realism*. Berkeley: University of California Press.
- _____. 1997. *A Novel Defense of Scientific Realism*. New York: Oxford University Press.

- Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. New York: Routledge.
- Putnam, Hilary (1978).
- Russell, Bertrand, 1931. *The Scientific Outlook*. London: George Allen & Unwin, Ltd.
- Saatsi, Juha (2005). "Reconsidering the Fresnel-Maxwell Case Study." *Studies in History and Philosophy of Science* 36 (3): 509–38.
- Saatsi, Juha and Peter Vickers (2011). "Miraculous Success? Inconsistency and Untruth in Kirchhoff's Diffraction Theory." *British Journal for the Philosophy of Science* 62: 29–46.
- Stanford, P. Kyle. 2006. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press.
- Teller, Paul (2015). "Language and the Complexity of the World;" forthcoming.
- Van Fraassen (1980). *The Scientific Image*. Oxford: Oxford University Press.
- Vickers, Peter (2013). "A Confrontation of Convergent Realism". *Philosophy of Science* 80: 189-211.
- Votsis, Ioannis (2011). "Saving the Intuitions: Polythetic Reference." *Synthese* 180 (2): 121–37.
- Worrall, J. 1989a. "Fix it and be Damned: A Reply to Laudan". *British Journal for the Philosophy of Science* (40): 376-388.
- (1989b). "Structural Realism: The Best of Both Worlds". *Dialectica* 43: 99-124.
- (2016). "Structural Realism – the Only Viable Realist Game in Town." Forthcoming in *Scientific Realism: Objectivity and Truth in Science*, Wenceslao Gonzalez & Evandro Agazzi, (eds.).

Concrete Models and Holistic Modelling*

Wei Fang[♠]

Department of Philosophy, University of Sydney

Abstract: This paper proposes a holistic approach to the model-world relationship, suggesting that the model-world relationship be viewed as an *overall structural fit* where one organized whole (the model) fits another organized whole (the target). This approach is largely motivated by the implausibility of Michael Weisberg's weighted feature-matching account of the model-world relationship, where a set-theoretic conception of the structures of models is assumed. To show the failure of Weisberg's account and the plausibility of my approach, a concrete model, i.e. the San Francisco Bay model, is discussed.

* Draft paper, please do not quote without permission.

[♠] Address: University of Sydney, NSW 2006, Australia. Email: wfan6702@uni.sydney.edu.au.

1. Introduction

One philosophical interest in the philosophy of modelling focuses on the problem of the model-world relationship, also known as the representation problem. Among many approaches to this problem, the similarity account has attracted much attention recently. Ronald Giere (1988, 1999a, 1999b, 2004, 2010), Peter Godfrey-Smith (2006) and Michael Weisberg (2012, 2013) have made the most substantial contributions.

The core of this account, first developed by Giere, is a view of the model-world relationship:

The appropriate relationship, I suggest, is *similarity*. Hypotheses, then, claim a *similarity* between models and real systems. But since anything is similar to anything else in some respects and to some degree, claims of similarity are vacuous without at least an implicit specification of relevant *respects and degrees*. The general form of a theoretical hypothesis is thus: Such-and-such identifiable real system is similar to a designated model in indicated respects and degrees. (Giere 1988, 81; author's emphasis)

However, critics point out that this account is only schematic since it falls short of specifying the relevant *respects and degrees* (Suárez 2003). Moreover, Giere argues that a philosophical account of scientific representation should also take into consideration factors such as the *roles* played by scientists, and the *intentions* those scientists have when modelling (Giere 2004, 2010). Given these considerations, Weisberg develops a more sophisticated similarity account, called the *weighted feature-matching* account

(2012, 2013). The basic idea of his account comes from psychologist Amos Tversky's *contrast* account of similarity, which states that the similarity of objects a and b depends on the features they share and the features they do not. In light of this, Weisberg proposes his own account:

S (m, t) =

$$\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m)$$

$$\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m) + \alpha f(M_a - T_a) + \beta f(M_m - T_m) + \gamma f(T_a - M_a) + \delta f(T_m - M_m) \quad (1)$$

$f(x)$ refers to the weighting function, α , β , γ , δ , θ , and ρ denote weighting terms (parameters), subscripts a and m stand for attributes and mechanisms,¹ and M denotes the model and T the target. $(M_a \cap T_a)$ stands for attributes shared by the model and the target, $(M_a - T_a)$ attributes that the model has while the target does not, and $(T_a - M_a)$ attributes that the target has while the model does not. The same story goes for mechanisms m . Attributes and mechanisms as a whole are called *features* of the model and the target.

An interpretation for this equation is needed. First, there must be a feature set \mathcal{A} , and the set of features of the model and the set of features of the target are defined as sets of features in \mathcal{A} . The elements of \mathcal{A} are determined by a combination of context, conceptualization of the target, and the theoretical goals of the scientist. Besides, the

¹ Properties and patterns of systems are termed attributes, and the underlying mechanisms generating these properties and patterns are termed mechanisms (Weisberg 2013, 145).

contents of \mathcal{A} may change through time as science develops, which in turn might result in a reevaluation of the established model-world relationship (*Ibid.*, 149).

Second, consider the values of weighting parameters α , β , γ , δ , θ , and ρ . On Weisberg's account, different kinds of modelling require different weighting parameters. For example, if what interests us is the *minimal modelling* which concerns merely the mechanism responsible for bringing about the phenomenon of interest, the goal of this modelling is written as:²

$$\frac{|M_m \cap T_m|}{|M_m \cap T_m| + |M_a - T_a| + |M_m - T_m|} \rightarrow 1 \quad (2)$$

Finally, consider the weighting function $f(x)$, telling us the relative importance of each feature in the set \mathcal{A} . Weisberg says scientists in most cases have in their mind some subset of the features in \mathcal{A} , which they regard as especially important. Hence some features are weighted more heavily, while others are equally weighted. Besides, the background theory determines which features in \mathcal{A} should be weighted more heavily. If the background theory is not rich enough, deciding which should be weighted more heavily is partly an empirical problem.

Having presented an outline of Weisberg's account, I will now argue that this account fails to capture the relationship between concrete models and their targets. To illustrate this

² Weisberg also describes three other kinds of modelling requiring different weighting parameters: hyperaccurate, how-possibly and mechanistic modelling (2013, 150-52).

shortcoming (Sec. 3), I will first describe the San Francisco Bay model (Sec. 2). Sec. 4 will propose a holistic alternative to Weisberg's account, suggesting that the model-world relationship be viewed as an *overall structural fit* where one organized whole fits another organized whole. Sec. 5 will examine a case where the organization of the whole can be treated as simply another feature.

2. The San Francisco Bay Model

John Reber worried about the fragility of the water supply in the San Francisco Bay area in the 1950s. To solve this problem, he proposed an ambitious proposal, namely, to dam up the Bay. Carrying out this plan would not only supply San Francisco with unlimited drinking water but also entirely change the area's transportation, industrial, military and recreation landscape (Weisberg 2013, 1). However, his critics worried that Reber's plan would only achieve its aims at the cost of destroying commercial fisheries, rendering the South Bay a brackish cesspool, creating problems for the ports of Oakland, Stockton, and Sacramento, and so on (Jackson and Peterson 1977; Cf. Weisberg 2013, 1).

To settle this dispute, the Army Corps of Engineers was charged with investigating the overall influence of the Reber plan by building a massive hydraulic scale model of the Bay (Weisberg 2013, 1-2). Once the model was built, it was adjusted to accurately reproduce several measurements of the parameters such as tide, salinity, and velocities actually recorded in the Bay (for details see Army Corps of Engineers 1963). After the adjustment, it was time to verify the model:

Agreement between model and prototype for the verification survey of 21-22 September 1956, and for other field surveys, was excellent. Tidal elevations, ranges and phases observed in the prototype were accurately reproduced in the model. Good reproduction of current velocities in the vertical, as well as in the cross section, was obtained at each of the 11 control stations in deep water and at 85 supplementary stations. The salinity verification tests for the verification survey demonstrated that for a fresh-water inflow into the Bay system [...], fluctuation of salinity with tidal action at the control points in the model was in agreement with the prototype (Huggins and Schultz 1967, 11).

After the verification, modellers were in a good position to assess the Reber plan through the model built. The investigation showed that it would considerably reduce water-surface areas, reduce the velocities of currents in most of South San Francisco Bay, reduce the tidal discharge through the Golden Gate during the tidal cycle, and so forth (Huggins and Schultz 1973, 19). Given these disastrous consequences, the Army Corps then denounced Reber's plan (Weisberg 2013, 9).

3. How Could Weisberg's Account Shed Light on the Bay Model?

I have argued elsewhere that Weisberg's account cannot shed light on mathematical models due to its atomistic conception of features and its assumption of the set-theoretic approach to model structures (citation anonymized). I find that the same charges can be raised in the case of concrete models.

Consider the first charge: Weisberg's account is committed to an atomic conception of features. The key of Weisberg's account is the claim that the similarity of objects a and b depends on the features they share and the features they do not share. Let us take a closer look at the equation (1). The numerator invites us to weight features shared, and the denominator asks us to weight all features involved (including three feature subsets: features shared, features possessed by the model but not the target, and features possessed by the target but not the model). Each feature is weighted independently and only once, with it falling into one of the three feature subsets. The numerator is the weighted sum of features shared, the denominator is the weighted sum of features shared and unshared, and the similarity measure is the ratio of the numerator to the denominator.

However, features in the Bay model are not atomistic and independent of each other. As Huggins and Schultz put it explicitly, "Among the problems to be considered were the conservation of water [...]; [...] the tides, currents and salinity of the Bay as they affect other problems [...]. None of these problems can be studied separately, for each affects the others" (1973, 12). The reason why none of these problems can be studied separately is because factors involved in these problems cannot be studied separately.

Consider, for instance, the relationship between two key features in the model: tide and salinity. Salinity levels vary along an estuary depending on the mixing of freshwater and saltwater at a site. An estuary "is the transition between a river and a sea. There are two main drivers: the river that discharges fresh water into the estuary and the sea that fills the estuary with salty water, on the rhythm of the tide" (Savenije 2005, Preface ix).

To illustrate this “rhythm of the tide”, consider the effect of the spring-neap tidal cycle on the vertical salinity structure of the James, York and Rappahannock Rivers, Virginia, U.S.A.:

Analysis of salinity data from the lower York and Rappahannock Rivers (Virginia, U.S.A.) for 1974 revealed that both of these estuaries oscillated between conditions of considerable vertical salinity stratification and homogeneity on a cycle that was closely correlated with the spring-neap tidal cycle, i.e. homogeneity was most highly developed about 4 days after sufficiently high spring tides while stratification was most highly developed during the intervening period. (Haas 1977, 485)

This short report shows not only that characteristics of salinity (such as stratification and homogeneity) are influenced by characteristics of the tide, but also that there is a phase connection (or synchronization) between tidal cycle and salinity oscillations. The former is a causal relationship while the latter is a temporal relationship. The phase connection among features was also emphasized by the Army Corps when verifying the Bay model, saying “These gages were installed in the prototype and placed in operation several months in advance of the date selected to collect the primary tidal current and salinity data required for model verification, since *it was essential to obtain all data simultaneously for a given tide over at least one complete tidal cycle of 24.8 hours*” (1963, 50; my emphasis).

Moreover, the same story goes for tide and tidal currents (for details see Army Corps 1963, 20).

In short, features in a model bear not only causal relationships, but also temporal relationships to one another. This implies that, when verifying the model, features of the

model causally interact with each other in producing certain outputs (e.g. predictions, effects, phenomena, etc.), rather than that they individually or separately produce outputs. So although outputs of key features in the Bay model can be identified and measured separately, they are not produced separately.

It is important to note that the causal interaction among features may lead to a different kind of interaction, i.e. a “similarity interaction”,³ wherein features interact with one another in producing the similarity value. That is, one feature’s contribution to the similarity value depends on other feature(s)’ contribution to that value.⁴ The difference between causal and similarity interaction is that the latter is a statistical relationship among measured features, and can be viewed as a reflection of the former when coupled with an assumption that there might be such an underlying causal structure.⁵ For example, a similarity interaction is shown by the verification of salinity in the Bay model, where the measurement of salinity (as a measurement of one feature’s contribution to the similarity

³ I thank X for suggesting this term for me.

⁴ This point can be best illustrated with the curve fitting example: when computing the fit of a straight line $y=ax+b$ to a cloud of points, a and b will depend on each other to produce the best fit (I thank X for giving me this example).

⁵ This assumption is important because there are cases where the fact that there is similarity interaction cannot guarantee that there is also causal interaction, because some randomly generated data set may also show interaction among features. In other words, causal interaction can lead to similarity interaction and the reverse is not true (I thank Y for letting me know this). I will discuss this assumption, called “precondition” later, in Sec. 4.

value from Weisberg's perspective) depended on other features in the way in which other features were kept constant: "salinity phenomena in the model were in agreement with those of the prototype *for similar conditions of tide, ocean salinity, and fresh-water inflow*" (*Ibid.*, 54; my emphasis).

The way that similarity interaction reflects causal interaction, when coupled with the assumption mentioned above, can be expressed as follows: if what is under verification is a causal structure to which modellers do not have direct access (so the structure cannot be a feature in Weisberg's formula), then the coherent behavior of features (i.e. their similarity interactions such as phase connections) is a way of verifying, or at least indicating, the causal interactions in the underlying causal structure.⁶ That is the reason why it was so essential to obtain all data simultaneously within a complete tidal cycle for the Bay model, and why all other features must be kept constant when verifying salinity (or other features).

Given features' causal interactions in the model and their similarity interactions when measuring them, it seems that assessing the relationship between a model and its target cannot be simply achieved in the way suggested by Weisberg's equation, for features' contribution to the similarity relationship is not *additive* but *interactive*. That is, to assess the relationship between a model and its target, one cannot measure each feature's contribution independently and then add them together.

4. Set-Theoretic or Non-Set-Theoretic? A Holistic Alternative

⁶ I thank X for bringing this point to my attention.

Now we arrive at the problem of why Weisberg's account is deeply committed to an atomistic conception of features. As I have argued elsewhere, this problem ultimately comes down to Weisberg's understanding of the structure of models (citation anonymized). Weisberg says models are *interpreted structures* (2013, 15), so concrete models are interpreted concrete structures. At first glance, I have no quarrel with this understanding. On closer inspection, however, it can be shown that Weisberg's account on the model-world relationship assumes a set-theoretic approach to the structure of models.⁷ This is because Weisberg's similarity measure can be derived from the *Jaccard similarity coefficient* between two sets, a coefficient assuming a set-theoretic conception of objects (citation anonymized).

The key to the set-theoretic approach to structures is its assumption that elements of objects (i.e. models and targets) are independent of each other, just as elements of a set are independent of each other. In other words, it construes both the model and the target as a set of independent elements, the similarity between which consists in the ratio of the number of elements shared to the number of all elements (citation anonymized). However, as discussed in Sec. 3, features are not independent. More importantly, their causal interactions may result in a similarity interaction among features.

This similarity interaction undermines Weisberg's account, for it cannot properly capture the dependence relationship of features' contribution to the overall similarity

⁷ Note that Weisberg *explicitly* objects to the set-theoretic approach to models (2013, 137-42). However, I think it is compatible to claim that someone *implicitly* assumes what someone explicitly rejects.

measure between a model and a target. Nonetheless, there is still a way to save the very intuitive notion of similarity, by abandoning the set-theoretic conception of structures. That is, if the structure of a model is viewed as an *organized whole* in which each component of the whole is interconnected to other component(s) (directly or indirectly) in such a way that they interact with one another in producing certain phenomena of interest (i.e. outputs). Under such an understanding, therefore, assessing the relationship between a model and its target cannot be simply achieved by assessing each individual feature's relationship and then adding them together. Nor can this be done by assessing each connection among two or more features and then adding them together, even if connections (causal or non-causal) are also interpreted as features. On the other hand, however, the notion of similarity can be minimally preserved by claiming that assessing the similarity or *fit* (I will use fit hereafter) between a model and a target amounts to assessing the *overall structural fit* between the model and its target.

Generally speaking, structural fit means the structure of the model fits the structure of the target *as an organized whole*. That said, nevertheless, it should be stressed that there is no univocal meaning for the term “structural fit” that could encompass all circumstances, nor can a single equation or formula capture all situations. This is largely due to the heterogeneity of modelling practice and its multifarious goals. On the other hand, however, instructive points can still be asserted. In what follows I will elaborate some basics regarding the conception of “structural fit”.

Structural fit in mathematical modelling means different things than in concrete modelling. For example, in a very simple case of curve fitting where a straight line $y=ax+b$

is fitted to a cloud of points, features a and b will interact with each other to produce the best fit. That is, what fits the cloud of points is the overall structure, not the additive sum of each individual feature. As I have argued elsewhere, in more complicated mathematical modelling such as the *maximum likelihood estimation*, the fit is usually achieved through comparing the predicted data set derived from the model *as a whole* to the observed data set derived from the target system (citation anonymized). Individual features of the model simply disappear, and causally related features, as constituting a whole, that co-occur in the data set are what really matters.

In the case of concrete modelling, admittedly, the claim that assessing the fit between a model and a target amounts to assessing the overall structural fit seems to be less apparent. On closer examination, however, the same claim still holds. Let us go back to the verification of the Bay model. At first glance, it seems the verification of the model was achieved by independently verifying the output (i.e. data sets) of each individual feature, as the report showed (see Sec. 2 for the verification report). That is, it seems that by verifying that each feature in the model fits its counterpart in the target, scientists made the judgment that the model fits the target system.

Underlying this seemingly plausible reasoning, however, there remains the problem of why we are allowed to confirm the verification of the model by means of only verifying several outputs of individual features. Or, to put it slightly differently, in terms of what does the fit of features guarantee the judgment about the fit of the model to the target? I take it that it is more than the fit of individual features themselves that makes sense of the reasoning that the model fits the target. There must be a precondition for this reasoning

(remember the “assumption” made in the last section). After all, there are many cases in which the fit of features does not guarantee the fit of the model itself to the target. For instance, a drawing of Tom’s face may accurately capture all features of his face, e.g., nose, eyes, mouth, etc., but still falls short of fitting his face, because of the wrong organization of these features, e.g., putting the mouth in between the eyes and nose (Weisberg would argue that the organization could be a feature. I will discuss this point in Sec. 5.).

So if the fit of features is insufficient to vindicate the fit of a model to its target, what could provide this vindication? My claim is, contrary to Weisberg, that it is the *overall structural fit* of the model to the target system that warrants the fit judgment about the model and its target. In other words, the fit of individual features can only succeed in supporting the fit of the model to the target by the precondition that these features can be organized into the whole (i.e. the assumption that there is such an underlying causal structure), not the other way around.

To understand this “holistic reasoning”, let me articulate the specifics involved step by step. We first build a concrete model, i.e. a concrete structure, wherein features are interconnected with one other in such a way that they have the potential to interactively produce certain phenomena of interest (i.e. outputs). Before verifying the model, we need to adjust key features to make sure the model works very well. Note that any adjustment will not simply be the adjustment of individual features but also of their interconnections, resulting in the adjustment of the overall structure of the model. Finally, we verify the model by comparing the outputs of the model to the outputs of the target. As with mathematical models, this verification is also usually made via comparing data sets, as

shown in the Bay model. Note that though these outputs can be identified, derived and measured independently, it is causally connected features that interact in producing them. In other words, although you verify each feature separately, the support provided by a single feature is not confined to that feature of the model, but confirms all aspects of the model that are involved in generating that output.

Thus understood, therefore, the gist of verifying a concrete model such as the Bay model can be captured as follows. The verification of each feature, as a component of a whole, is simply the verification of one aspect of the structure. So the verification of different features is the verification of the same structure from different perspectives. Thus, if the model is an organized whole, then the more features that are independently verified the more likely it is that the model resembles the reality. On the other hand, if what is under verification is not an organized whole but an aggregation of independent items, then the verification of each lends no credence to other parts of the aggregated whole—because these items are not causally linked, the verification of each item is only the verification of that item itself.

In sum, the relationship between a concrete model and its target is a holistic matter wherein an organized whole fits (to a certain degree) or fails to fit another organized whole. Though it seems at first blush that the verification of the whole results from the sum of the verification of each component, the real picture is just the reverse: the whole is always in place and the component can gather force in supporting the verification of the whole only when it can be organized into the whole.

5. Organization and Features

As mentioned above, Weisberg would argue that the organization could be a feature, so a drawing of Tom's face capturing accurately not only his nose, mouth, eyes but also their organization can be a good model of Tom's face. A holistic account agrees that organization could be a feature, but disagrees with the way that organization is treated in Weisberg's similarity measure. Intuitively, we may say that a drawing of one person's face is a good model if it has the right features: such as a nose, a mouth, eyes, and the organization of all of these. So it seems that if you get each individual feature right, then you get the whole model right. That is, features *additively* contribute to the goodness of the model.

This intuitive way of understanding scientific modelling, however, obscures the fact that features may interact in producing the fit of a model, as shown in Sec. 4. To reiterate this point and to draw a connection to our current discussion, consider another ordinary example.⁸ Suppose Anne's face is an ideal one which scientists want to model. Anne has an ideal nose, which is straight, in contrast to a non-ideal nose, which might be bumped or concave. She also has an ideal nostril, which is round, in contrast to a non-ideal one, which might be triangular or square. Scientist A draws a face for Anne that has a round nostril and a concave nose, while scientist B draws a face that has a triangular nostril and a bumped nose. Drawing A has an ideal feature (the round nostril), but neither feature of drawing B is ideal. Now we ask which drawing better fits Anne's face. It is likely that we

⁸ I thank X for giving me this nice example.

will say that B is better because our contemporaries' taste tells us that there is no face so ugly as one with a round nostril and a concave nose, though a round nostril itself is ideal. Hence we see a case wherein the nostril and nose interact to produce the fit of a model to a target.

This discussion leads to a more general question: what are features? In Weisberg's account, a model can *more or less* fit a target, but features are either shared or not. Yet as Wendy Parker points out, "relevant similarities often seem to occur at the level of individual features, not just at the level of the model" (2015, 273). This is because features themselves can be objects such that they more or less fit each other.⁹ Weisberg may argue that this problem can be fixed by the assumption that a feature can be redescribed as a set of sub-features, so the similarity between two features can be measured as the result of the similarity between their sub-features. However, I see this treatment as a non-starter, for the similarity between sub-features may also be a matter of degree such that it should be measured as the result of the similarity between their sub-sub-features, and between their sub-sub-sub-features, and so on.

On the other hand, a holistic account does not encounter this problem: if a feature is an object, then it can be viewed as an organized whole. So the relationship between a feature in a model and a feature in a target also consists in their structural fit. Take a minimal model for instance. Most minimal models primarily attempt to represent repeatable patterns of behavior largely insensitive to underlying microscopic details (Batterman 2002, 27). Suppose we are interested in the buckling behavior of struts, and write a

⁹ I thank X for bringing this to my attention.

phenomenological formula, called Euler's formula, to characterize it (see Batterman 2002 for details). It seems the pattern of behavior is the only feature involved in this case, i.e., a dependence relationship among several parameters. So assessing the fit between the model and the target comes down to assessing the fit between the feature in the model and the feature in the target. For this, a holistic account can easily come through: the relationship is an overall structural fit, wherein a dependence relationship as a feature fits another dependence relationship.

6. Conclusion

This paper has shown that the assumption of a set-theoretic approach to structures makes Weisberg's account fail to shed light on the San Francisco Bay model. Alternatively, a holistic approach to models, viewing the model-world relationship as an overall structural fit, fares better not only in capturing the Bay model, but more generally in making sense of modelling practice.

References

- Army Corps of Engineers. 1963. *Technical Report on Barriers: A Part of the Comprehensive Survey of San Francisco Bay and Tributaries, California*. Appendix H, Volume 1: Hydraulic Model Studies. San Francisco: Army Corps of Engineers.
- Batterman, Robert. 2002. "Asymptotics and the Role of Minimal Models." *British Journal for the Philosophy of Science* 53 (1): 21-38.
- Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.
- Giere, Ronald N. 1999a. *Science without Laws*. Chicago: University of Chicago Press.
- Giere, Ronald N. 1999b. "Using Models to Represent Reality." In *Model-Based Reasoning in Scientific Discovery*, ed. Lorenzo Magnani, Nancy J. Nersessian, and Paul Thagard, 41-57. Springer Science & Business Media.
- Giere, Ronald N. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71 (5): 742-752.
- Giere, Ronald N. 2010. "An Agent-Based Conception of Models and Scientific Representation." *Synthese* 172 (2): 269-281.
- Godfrey-Smith, Peter. 2006. "The Strategy of Model-based Science." *Biology and Philosophy* 21 (5): 725-740.
- Haas, Leonard W. 1977. "The Effect of the Spring-Neap Tidal Cycle on the Vertical Salinity Structure of the James, York and Rappahannock Rivers, Virginia, U.S.A." *Estuarine and Coastal Marine Science* 5:485-496.

- Huggins, Eugene M., and Edward A. Schultz. 1967. "San Francisco Bay in A Warehouse." *Journal of the IEST* 10 (5): 9-16.
- Huggins, Eugene M., and Edward A. Schultz. 1973. "The San Francisco Bay and the Delta Model." *California Engineer* 51 (3): 11-23.
- Jackson, W. Turrentine, and Alan M. Peterson. 1977. *The Sacramento-San Joaquin Delta: The Evolution and Implementation of Water Policy*. Davis: California Water Resource Center, University of California.
- Parker, Wendy. 2015. "Getting (even more) serious about similarity." *Biology and Philosophy* 30 (2): 267-276.
- Savenije, Hubert H. G. 2005. "Salinity and Tides in Alluvial Estuaries." Elsevier Science.
- Suárez, Mauricio. 2003. "Scientific Representation: against Similarity and Isomorphism." *International Studies in the Philosophy of Science* 17 (3): 225-244.
- Weisberg, Michael. 2012. "Getting Serious about Similarity." *Philosophy of Science* 79 (5): 785-794.
- Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

PROBABILISTIC ACTUAL CAUSATION

LUKE FENTON-GLYNN

DEPARTMENT OF PHILOSOPHY, UNIVERSITY COLLEGE LONDON

GOWER STREET, LONDON, WC1E 6BT, U.K.

ABSTRACT. Actual (token) causes – e.g. Suzy’s being exposed to asbestos – often bring about their effects – e.g. Suzy’s suffering mesothelioma – probabilistically. I use probabilistic causal models to tackle one of the thornier difficulties for traditional accounts of probabilistic actual causation: namely probabilistic preemption.

Luke Fenton-Glynn

1. INTRODUCTION

Actual (token) causation is the relation that obtains when, for example, Suzy's being exposed to asbestos causes her to suffer mesothelioma. A number of theorists (e.g. Halpern and Pearl 2001, 2005; Hitchcock 2001, 2007; Weslake 2016) have deployed structural equations models (SEMs) in developing novel solutions to difficulties confronting traditional accounts of this relation. These theorists have focused on *deterministic* actual causation (DAC).¹ I draw on probabilistic causal models (PCMs) – analogues of deterministic SEMs – to provide an account of probabilistic actual causation (PAC). I don't attempt to show that my account can handle the full battery of test cases discussed in the literature. I simply demonstrate that it yields an elegant treatment of one very central case – probabilistic preemption – with a view to motivating further investigation of formal approaches to PAC.

2. PROBABILITY-RAISING

Probability-raising is central to the account developed here – as on traditional accounts of PAC.² To explain how I will understand that notion a bit of stage-setting is required.

I take the relata of the actual causal relation to be variable values. Adopting Goldszmidt and Pearl's (1992, 669–70) notation, $P(W = w|do(V = v))$ represents the probability for $W = w$ that *would* obtain if V were set to $V = v$ by an 'intervention' (Woodward 2005, 98). This is liable to diverge from the conditional probability $P(W = w|V = v)$: witness the difference between the probability of a storm *conditional* upon the barometer needle pointing toward the

¹Cf. Halpern and Pearl (2005, 852); Hitchcock (2007, 498).

²Reichenbach (1971, 204); Suppes (1970); Lewis (1986, 175–84); Menzies (1989). The deficiencies of these accounts have been demonstrated by e.g. Salmon (1984, 192–202); Menzies (1996, 85–96); Hitchcock (2004).

Probabilistic Actual Causation

word ‘storm’ and the probability of a storm if I had intervened upon the barometer needle to point it toward ‘storm’.

Variable X taking value $X = x$ (rather than $X = x'$) raises the probability of $Y = y$ in the relevant sense iff:³

$$(1) \quad P(Y = y|do(X = x)) > P(Y = y|do(X = x'))$$

Appealing to interventionist probabilities means avoiding probability-raising relations between independent effects of a common cause, such as the barometer reading and the storm (cf. Lewis 1986, 178).

Probabilistic preemption cases illustrate that straightforward probability-raising is neither necessary nor sufficient for causation (Menzies 1989, 1996).

3. PROBABILISTIC PREEMPTION

The following example is inspired by Anscombe (1971).⁴

³Here and throughout, the probabilities (chances) should be taken to be those obtaining immediately after the interventions bringing about the variable values specified in the scope of the $do(\cdot)$ function have occurred (cf. Lewis 1986, 177).

⁴The probabilities involved (except the decision probabilities) are quantum and therefore objective and able underwrite causal relations. (If you’re worried that the decision probabilities are not objective, the example could be complicated so that the decisions are made on the basis of outcomes of quantum measurements.) I find it plausible that the probabilities of many high level sciences are also objective (cf. e.g. Loewer 2001; Ismael 2009).

Luke Fenton-Glynn

(ProbPre) *Someone (neither you nor I) has connected a Geiger counter to a bomb so that the bomb will explode if the Geiger registers above a threshold reading. I place a place a chunk of U-232 (half-life = 68.9 years; decays by α -emission) near the Geiger. By chance, enough U-232 atoms decay within a short enough interval for the Geiger to reach the threshold reading so that the bomb explodes. Unbeknownst to me, you've been standing nearby observing. You have a chunk of Th-228 (half-life = 1.9 years; decays by α -emission), which contains many more atoms than my chunk of U-232. You've decided that you'll place your Th-228 near the Geiger iff I fail to place my U-232 near the Geiger. There's a negligible chance that you won't follow the course of action you've decided on. Seeing that I place my U-232 near the Geiger, you don't place your Th-228 near the Geiger.⁵*

Let M , D , Y , T , and E be binary variables which, respectively, take value 1 if the following things occur (and 0 otherwise): I place my U-232 near the Geiger; you decide to place your Th-228 near the Geiger iff I don't place my U-232 near the Geiger; you place your Th-228 near the Geiger; the threshold reading is reached; the bomb explodes.

My act ($M = 1$) was an actual cause of the explosion ($E = 1$). Yet plausibly the following inequality holds:

$$(2) \quad P(E = 1 | do(M = 1)) < P(E = 1 | do(M = 0))$$

⁵The range of α -particles is 3-5 cm. Suppose that, for each of us, a decision to place our chunk 'near' the Geiger counter is a decision to place it < 5 cm away and a decision not to place it nearby is a decision to place it nowhere near ($\gg 5$ cm away).

Probabilistic Actual Causation

That is, my placing my U-232 near the Geiger *lowers* the probability of the bomb exploding because it strongly lowers the probability of your placing your more potent Th-228 near the Geiger. Probability-raising is therefore unnecessary for actual causation.

Your decision ($D = 1$) was *not* an actual cause of the explosion, since you don't place your Th-228 near the Geiger. Yet provided there's some chance that $M = 0$, the following inequality holds:

$$(3) \quad P(E = 1 | do(D = 1)) > P(E = 1 | do(D = 0))$$

Inequality (3) holds because your decision raises the probability that the bomb will still explode in the scenario in which $M = 0$.⁶ Probability-raising is therefore insufficient for actual causation.

Actual causation therefore can't be identified with probability-raising. In developing a more nuanced analysis, it is helpful to appeal to PCMs.

4. PCMs

A PCM, \mathcal{M} , is a 5-tuple $\langle \mathcal{V}, \mathcal{C}, \Omega, \mathcal{F}, do(\cdot) \rangle$. \mathcal{V} is a set of variables. Suppose \mathcal{R} denotes a function from elements of \mathcal{V} to sets of values: for all $V \in \mathcal{V}$, $\mathcal{R}(V)$ is the *range* of V . In Halpern and Pearl's (2005, 851–2) terminology, a formula $V_i = v_i$, for $V_i \in \mathcal{V}$ and $v_i \in \mathcal{R}(V)$, is a *primitive event*. \mathcal{C} is the set of all those possible conjunctions of primitive

⁶ $D = 0$ is multiply realizable: there is more than one alternative to the decision that you in fact make. E.g. you could decide that you will place your Th-228 near the Geiger no matter what, or that you will not do so no matter what. We can stipulate that the latter alternative is much more probable.

Luke Fenton-Glynn

events, $V_1 = v_1 \& \dots \& V_n = v_n$, such that $V_i \in \mathcal{V}$ and $v_i \in \mathcal{R}(V_i)$ and such that, for no pair of conjuncts $V_i = v_i, V_j = v_j$ is $V_i \equiv V_j$, and where no two elements of \mathcal{C} differ *only* in the permutation of their conjuncts. Such a conjunction is denoted $\mathbf{V} = \mathbf{v}$ (primitive events and the null event are limiting cases of such conjunctions). Abusing notation, the fact that $v_i \in \mathcal{R}(V_i)$ for each primitive event $V_i = v_i$ in the conjunction $\mathbf{V} = \mathbf{v}$, is abbreviated $\mathbf{v} \in \mathcal{R}(\mathbf{V})$ and the set of variables that appear in $\mathbf{V} = \mathbf{v}$ is denoted \mathbf{V} .

Call a conjunction $\mathbf{V} = \mathbf{v}$ *maximal* if it contains a conjunct of the form $V_i = v_i$ for each $V_i \in \mathcal{V}$. Ω is the set of all maximal conjunctions of primitive events. \mathcal{F} is a sigma algebra on Ω . Finally, $do(\cdot)$ is a function from elements of \mathcal{C} to probability distributions on \mathcal{F} (cf. Pearl 2009, 70, 110): for each element $\mathbf{V} = \mathbf{v}$ of \mathcal{C} , $P(\cdot | do(\mathbf{V} = \mathbf{v}))$ is the probability (chance) distribution on \mathcal{F} that *would* obtain if interventions were performed to bring about $\mathbf{V} = \mathbf{v}$.

A PCM can be represented graphically by taking the variables in \mathcal{V} as nodes and drawing a directed edge from V_i to V_j ($V_i, V_j \in \mathcal{V}$) iff, where $\mathbf{S} = \mathcal{V} \setminus V_i, V_j$, there is some assignment of values $\mathbf{s}' \in \mathcal{R}(\mathbf{S})$, some pair of values $v_i, v'_i \in \mathcal{R}(V_i)$ ($v_i \neq v'_i$) and some value $v_j \in \mathcal{R}(V_j)$ such that $P(V_j = v_j | do(V_i = v_i \& \mathbf{S} = \mathbf{s}')) \neq P(V_j = v_j | do(V_i = v'_i \& \mathbf{S} = \mathbf{s}'))$.

In constructing a PCM, \mathcal{M}_{Pre} , of **(ProbPre)** we might take the variable set to be $\mathcal{V}_{Pre} = \{D, M, Y, T, E\}$. The range of each variable in \mathcal{V}_{Pre} is the pair $\{0, 1\}$. \mathcal{C}_{Pre} , Ω_{Pre} , and \mathcal{F}_{Pre} are generated by \mathcal{V}_{Pre} and \mathcal{R}_{Pre} in the way described above. For each element of \mathcal{C}_{Pre} , the function $do(\cdot)$ returns the chance distribution on \mathcal{F}_{Pre} that would obtain if interventions were performed to bring about that element of \mathcal{C}_{Pre} . The graph for \mathcal{M}_{Pre} is given as figure 1.

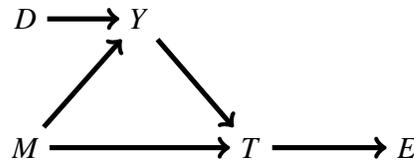


FIGURE 1

Probabilistic Actual Causation

A directed path in a graph is an ordered sequence of nodes, $\langle V_1, V_2, \dots, V_n \rangle$, such that there is a directed edge from V_1 to V_2 , and a directed edge from V_2 to $\dots V_n$. $\langle M, Y, T, E \rangle$ is an example of a directed path in the graph of \mathcal{M}_{Pre} .

5. APPROPRIATE MODELS

In Section 6, I provide a definition of what it is for $X = x$ (rather than $X = x'$) to count as an actual cause of $Y = y$ *relative to a PCM*. I then define a non-model-relativized notion of actual causation by saying that $X = x$ (rather than $X = x'$) counts as an actual cause of $Y = y$ *simpliciter* provided that $X = x$ (rather than $X = x'$) counts as an actual cause $Y = y$ relative to at least one *appropriate* PCM.⁷ A similar strategy is commonly adopted by those analyzing DAC in terms of SEMs (Hitchcock 2001, 287, 2007, 503; Weslake 2016). This requires an account of ‘appropriate’ models.

Many of the criteria for an appropriate SEM for evaluating DAC carry over to PCMs, including the following three:

(Partition) For all $V \in \mathcal{V}$, the elements of $\mathcal{R}(V)$ should form a partition (Halpern and Hitchcock 2010, 397–8; Blanchard and Schaffer 2016)

(Independence) For no two variables $V, W \in \mathcal{V}$ should there be elements $v \in \mathcal{R}(V)$ and $w \in \mathcal{R}(W)$ such that the states of affairs represented by $V = v$ and $W = w$ are logically or metaphysically related (Hitchcock 2001, 287; Halpern and Hitchcock 2010, 397)

⁷As the parentheses indicate I define a *contrastive* relation of actual causation. Where variables are binary – as in \mathcal{M}_{Pre} – this is inconsequential and I will typically suppress such parentheses. But it becomes important in cases of multi-valued variables (see Halpern and Pearl 2005, 859).

Luke Fenton-Glynn

(Naturalness) For all $V \in \mathcal{V}$, $\mathcal{R}(V)$ should include only values that represent reasonably natural and intrinsic states of affairs. (Blanchard and Schaffer 2016)

The analysis of actual causation proposed below takes all and only values of distinct variables to be potential causal relata. (Partition) insures that we don't thereby miss actual causal relations because they obtain between the values of a single variable. (Independence) insures that we don't mistake stronger-than-causal relations for causal relations. (Naturalness) insures that unnatural or non-intrinsic states of affairs do not get counted as causes and effects (see Lewis 1986, 190, 263; Paul 2000, 245).⁸

A further condition is that a model is appropriate for evaluating whether $X = x$ is an actual cause of $Y = y$ in world θ only if it satisfies (Veridicality):

(Veridicality) For any conjunction $\mathbf{V} = \mathbf{v} \in \mathcal{C}$ taken as an input, the probability distribution $P(\cdot | do(\mathbf{V} = \mathbf{v}))$ yielded as an output by $do(\cdot)$ should be the *objective chance* distribution over \mathcal{F} that would $_{\theta}$ result from interventions setting $\mathbf{V} = \mathbf{v}$. ('Would $_{\theta}$ ' indicates that what is required is that this counterfactual be true in θ .)

(Veridicality) is an analogue – for PCMs – of the requirement that SEMs encode only true counterfactuals (Hitchcock 2001, 287, 2007, 503).

In the DAC/SEMs literature another condition on model appropriateness is typically added:

(Serious Possibilities) \mathcal{V} should not be such as to generate elements of Ω that represent possibilities “that we consider to be too remote” (Hitchcock 2001, 287;

⁸If *absences* are unnatural states of affairs (cf. Lewis 1986, 189–93), we might instead require that each variable have *at most one value* representing such a state of affairs.

Probabilistic Actual Causation

cf. Woodward 2005, 86–91, Weslake 2016, Blanchard and Schaffer 2016).

We likely need this requirement too. A discussion of whether the vagueness and subjectivity thereby introduced is problematic would take us too far afield.⁹ Still, it doesn't put the present account in any *worse* shape than its deterministic analogues. Moreover, traditional accounts of actual causation – which don't appeal to causal models – also stand in need of appeal to 'serious possibilities' (Woodward 2005, 86–8).

A final requirement – similar to one imposed in the DAC/SEM literature – for a model \mathcal{M} to be an appropriate one for evaluating whether $X = x$ is an actual cause of $Y = y$ in world θ is:

(Stability) There is no model \mathcal{M}^* (satisfying Partition, Independence, Naturalness, Veridicality, and Serious Possibilities) with a variable set \mathcal{V}^* such that $\mathcal{V}^* \supset \mathcal{V}$ relative to which $X = x$ (rather than $X = x'$) is *not* an actual cause of $Y = y$. (Halpern and Hitchcock 2010, 394–5; Blanchard and Schaffer 2016; Halpern 2014; Hitchcock 2007, 503).

The idea is that an appropriate model is a sufficiently rich representation of causal reality that moving to a richer representation would not reveal an apparent actual causal relation to be spurious.¹⁰

The converse requirement – that a negative verdict about actual causation should not be overturned in a richer model – isn't needed. This is because actual causation (simpliciter) is defined in terms of actual causation relative to *at least one* appropriate model. A model relative verdict that $X = x$ is not an actual cause of $Y = y$ thus automatically fails to translate

⁹See Woodward (2005, 86–91).

¹⁰(Stability) renders the notion of an appropriate model relative to the causal claim being evaluated.

Luke Fenton-Glynn

into a verdict that $X = x$ is not an actual cause (simpliciter) of $Y = y$ if there is a richer (and otherwise appropriate) model relative to which $X = x$ is an actual cause of $Y = y$.

We can now state a definition of actual causation in terms of appropriate PCMs that handles **(ProbPre)**.

6. PAC

Actual causation *simpliciter* is defined in terms of actual causation relative to an appropriate PCM. Model-relative actual causation is then defined.¹¹

AC(S)

Where $x, x' \in \mathcal{R}(X)$ and $y \in \mathcal{R}(Y)$, $X = x$ (rather than $X = x'$) is an actual cause (simpliciter) of $Y = y$ in world θ iff $X = x$ (rather than $X = x'$) is an actual cause of $Y = y$ relative to at least one model \mathcal{M} (with $X, Y \in \mathcal{V}$) that is appropriate for evaluating whether $X = x$ (rather than $X = x'$) is an actual cause (simpliciter) of $Y = y$ in θ .

¹¹Those familiar with Halpern and Pearl's (2001, 2005) analyses of DAC are invited to see an analogy with **AC(M-R)**. **AC(M-R)** was partly inspired by thinking about how a counterpart of Halpern and Pearl's analysis might be developed that is adequate to the probabilistic case. Ultimately, I'm optimistic that an adequate account of DAC will fall out of an adequate account of PAC as the special case where all probabilities are 1 or 0. This is why my definitions take the definiendum to be 'actual cause' rather than 'probabilistic actual cause'.

Probabilistic Actual Causation

AC(M-R)

Where $x, x' \in \mathcal{R}(X)$ and $y \in \mathcal{R}(Y)$, $X = x$ (rather than $X = x'$) is an *actual cause* of $Y = y$ relative to a model \mathcal{M} (with $X, Y \in \mathcal{V}$) in world θ iff there is a partition (\mathbf{Z}, \mathbf{W}) of $\mathcal{V} \setminus X, Y$ and some setting $\mathbf{W} = \mathbf{w}'$ of the variables in \mathbf{W} such that the $do(\cdot)$ function associated with \mathcal{M} entails that, for all subsets \mathbf{Z}' of \mathbf{Z} (where, for each such subset, $\mathbf{Z}' = \mathbf{z}^*$ are the values that the variables in \mathbf{Z}' have in θ):

$$(IN) P(Y = y | do(X = x \& \mathbf{W} = \mathbf{w}' \& \mathbf{Z}' = \mathbf{z}^*)) > P(Y = y | do(X = x' \& \mathbf{W} = \mathbf{w}'))$$

AC(M-R) counts $M = 1$ as an actual cause of $E = 1$ relative to \mathcal{M}_{Pre} (and the world described in **(ProbPre)**). Consider the partition of $\mathcal{V}_{Pre} \setminus M, E$ such that $\mathbf{W} = \{D, Y\}$ and $\mathbf{Z} = \{T\}$. And consider the assignment $\{D = 1, Y = 0\}$ of values to the variables in \mathbf{W} . **AC(M-R)** is satisfied because **(IN)** holds for both subsets of \mathbf{Z} (\emptyset and $\{T\}$), as shown by (4) and (5):

$$(4) \quad P(E = 1 | do(M = 1 \& D = 1 \& Y = 0)) > P(E = 1 | do(M = 0 \& D = 1 \& Y = 0))$$

$$(5) \quad P(E = 1 | do(M = 1 \& T = 1 \& D = 1 \& Y = 0)) > P(E = 1 | do(M = 0 \& D = 1 \& Y = 0))$$

Inequality (4) indicates that my action raises the probability of the explosion *under the contingency* – i.e. *holding fixed* – that (you make your decision but) don't place your Th-228 near the Geiger. The existence of this *contingent* probability-raising reflects the fact that there is a path – $\langle M, T, E \rangle$ – along which $M = 1$ promotes $E = 1$ (because $M = 1$ raises the probability of $E = 1$ when we hold fixed the values of all variables off that path). It is the existence of

Luke Fenton-Glynn

such a path – representing the process via which $M = 1$ produces $E = 1$ – that appears to drive our intuitions about actual causation in this case (cf. Hitchcock 2001).

Inequality (5) indicates that, again holding fixed $D = 1$ and $Y = 0$, the probability of $E = 1$ is higher if I place my U-232 near the Geiger *and the threshold reading is reached* than if I'd simply never placed my U-232 near the Geiger in the first place. As will be seen, this requirement ensures that, not only is there a potential process via which $M = 1$ threatens to bring about $E = 1$, but that process is complete.

Since **AC(M-R)** implies that $M = 1$ is an actual cause of $E = 1$ relative to \mathcal{M}_{Pre} , **AC(S)** yields the (correct) result that $M = 1$ is an actual cause (simpliciter) of $E = 1$ provided that \mathcal{M}_{Pre} is appropriate. \mathcal{M}_{Pre} is appropriate. Clearly it satisfies (Partition) and (Independence). It satisfies (Naturalness) because all of the states that its variables represent are reasonably natural. It was stipulated that the $do(\cdot)$ function associated with \mathcal{M}_{Pre} is such that (Veridicality) is satisfied. \mathcal{M}_{Pre} does not represent the sort of ‘non-serious’ possibility that (Serious Possibilities) is introduced to rule out (cf. Hitchcock 2001; Woodward 2005, 86–91).

Finally, (Stability) is satisfied because the causal process from my action to the explosion is complete. Holding fixed $Y = 0$, the probability of the explosion if $M = 1$ *and* part(s) of this process occur(s) is higher than the probability of the explosion if simply $M = 0$. Any variable (whose values represent reasonably natural states, form a partition, and are logically and metaphysically independent from the variables in \mathcal{V}_{Pre}) that might be added to \mathcal{V}_{Pre} either represents part of this process or it doesn't. If it does, its actual value represents *the occurrence* of part of the process. So, if it is added to \mathcal{V}_{Pre} , including it in \mathbf{Z} will not prevent **(IN)** from holding for all subsets \mathbf{Z}' of \mathbf{Z} . If it doesn't, then adding it to \mathcal{V}_{Pre} , including it in \mathbf{W} , and holding it fixed at its actual value as part of the assignment $\mathbf{W} = \mathbf{w}'$ will not make a difference to the fact that **(IN)** holds for all subsets \mathbf{Z}' of \mathbf{Z} , since holding fixed $Y = 0$ as part of $\mathbf{W} = \mathbf{w}'$ is already sufficient to ensure this.

Probabilistic Actual Causation

AC(M-R) gives the verdict that $D = 1$ is *not* an actual cause of $E = 1$ relative to \mathcal{M}_{Pre} .

Consider the partition of $\mathcal{V}_{Pre} \setminus D, E$ such that $\mathbf{W} = \{M\}$ and $\mathbf{Z} = \{Y, T\}$. Observe that:

$$(6) \quad P(E = 1 | do(D = 1 \& M = 0)) > P(E = 1 | do(D = 0 \& M = 0))$$

And:

$$(7) \quad P(E = 1 | do(D = 1 \& M = 1)) > P(E = 1 | do(D = 0 \& M = 1))$$

Thus, whichever possible value we hold fixed M at, the probability of $E = 1$ is higher if $D = 1$ than if $D = 0$. So $D = 1$ contingently raises the probability of $E = 1$.¹² That's because there's a path – $\langle D, Y, E \rangle$ – along which $D = 1$ promotes $E = 1$.

AC(M-R) nevertheless entails that $D = 1$ is *not* an actual cause of $E = 1$ relative to \mathcal{M}_{Pre} . Consider the subset $\{Y\}$ of \mathbf{Z} , and observe that:

$$(8) \quad P(E = 1 | do(D = 1 \& Y = 0 \& M = 0)) \leq P(E = 1 | do(D = 0 \& M = 0))$$

And:

$$(9) \quad P(E = 1 | do(D = 1 \& Y = 0 \& M = 1)) \leq P(E = 1 | do(D = 0 \& M = 1))$$

That is, whichever possible value we hold fixed M at, the probability of the explosion is no higher if you make your decision *but don't place your Th-228 near the Geiger* than if you'd

¹²The obtaining of just one of (6) or (7) would suffice to show this.

Luke Fenton-Glynn

never made that decision in the first place. Thus **(IN)** does not hold for every subset of **Z** for this partition of variables no matter what values we assign to the variables in **W**. This reflects the fact that, because you didn't place your Th-228 near the Geiger, there is no complete causal process by which your decision produces the explosion. Your non-placement of your Th-228 'neutralizes' the danger of your decision causing the explosion.

Is there an alternative partition **(W, Z)** of \mathcal{V}_{Pre} and assignment **W** = **w'** such that **(IN)** holds for all subsets **Z'** of **Z**? (There need only be *one* for **AC(M-R)** to be satisfied.) There isn't. Assigning *Y* to **W** instead of **Z** won't help, since the value of *Y* 'screens off' *D* from *E*. So, where *Y* ∈ **W**, no assignment **W** = **w'** will be such that, holding fixed **W** = **w'**, the probability of *E* = 1 is higher when *D* = 1 (and the variables in $\emptyset \subseteq \mathbf{Z}$ are set to their actual values) than when *D* = 0. So **(IN)** doesn't hold for all subsets **Z'** of **Z** for any such partition.

On the other hand, if we leave *Y* in **Z** and also assign *M* to **Z**, then there are no variables in **W** to hold fixed. Now consider the subset {*Y*} of **Z**, and observe that:¹³

$$(10) \quad P(E = 1 | do(D = 1 \& Y = 0)) \leq P(E = 1 | do(D = 0))$$

So, with *M* assigned to **Z** it remains the case that **(IN)** doesn't hold for all subsets of **Z**.

So there's no partition of $\mathcal{V}_{Pre} \setminus D, E$ such that **(IN)** is satisfied for all subsets of **Z** when we consider *D* = 1 as a putative cause of *E* = 1. **AC(M-R)** therefore doesn't count *D* = 1 as an actual cause of *E* = 1 relative to \mathcal{M}_{Pre} .

But for **AC(S)** to count *D* = 1 as an actual cause of *E* = 1 *simpliciter*, there need only be one appropriate model relative to which **AC(M-R)** counts *D* = 1 as an actual cause of *E* = 1. Is there such a model? There isn't. Suppose a candidate such model includes *Y*. Because *D* is only relevant to *E* because of its relevance to *Y*, the value of *Y* 'screens off' the value of *D*

¹³Note: the fact that *Y* = 0 *due to an intervention* doesn't make *M* = 1 more likely.

Probabilistic Actual Causation

from that of E . This means that, if Y is included in \mathbf{W} in the partition (\mathbf{W}, \mathbf{Z}) of the model's variable set and held fixed (either at 1 or 0) as part of the assignment $\mathbf{W} = \mathbf{w}'$, then (\mathbf{IN}) won't be satisfied for the empty subset of \mathbf{Z} . Alternatively, if Y is included in \mathbf{Z} then, no matter what other variables are included in the model and assigned to \mathbf{W} , (\mathbf{IN}) won't be satisfied for the subset $\{Y\}$ of \mathbf{Z} . Specifically, because $D = 1$ only threatens to bring about $E = 1$ because it threatens to bring about $Y = 1$, no matter what we hold fixed by inclusion on both sides of (\mathbf{IN}) , the probability of $E = 1$ is no higher if $D = 1$ and $Y = 0$ than if simply $D = 0$.

So $\mathbf{AC}(\mathbf{M-R})$ doesn't count $D = 1$ as an actual cause of $E = 1$ relative to any appropriate model with Y in its variable set. This means that any otherwise appropriate model relative to which $D = 1$ is an actual cause of $E = 1$ can be expanded to a model in which $D = 1$ isn't an actual cause of $E = 1$ simply by the addition of Y . Provided the expanded model is appropriate, the original model violates (Stability) and is inappropriate. So $\mathbf{AC}(\mathbf{S})$ will correctly not count $D = 1$ as an actual cause *simpliciter* of $E = 1$.

Since the values of Y form a partition and represent natural states of affairs, (Partition) and (Naturalness) will be satisfied by the expanded model if they were satisfied by the original model. With regard to (Veridicality), it should be noted that there are multiple ways of expanding the original model via the addition of Y , each associated with a different $do(\cdot)$ function from elements of \mathcal{C}^* to probability distributions over \mathcal{F}^* (where \mathcal{C}^* and \mathcal{F}^* are generated by the expanded variable set in the way described in Section 4). In looking for an apt expanded model, we just select the one with the $do(\cdot)$ function that returns the objective chances on \mathcal{F}^* that *would* obtain as a result of interventions bringing about the various elements of \mathcal{C}^* . With regard to (Serious Possibilities) note that, given your decision, your placing *and* your not placing your Th-228 near the Geiger are both salient possibilities in

Luke Fenton-Glynn

(ProbPre). So it doesn't seem that the expanded model could represent any non-serious possibilities if the original model doesn't. (Independence) is a little trickier. Might not the original model include a variable whose values are logically or metaphysically related to those of Y ? Given that the variables in the original model are assumed to satisfy (Partition) it seems that any variable logically or metaphysically related to Y – e.g. Y' , which takes value $Y' = 0$ if you don't place your Th-228 near the Geiger, $Y' = 1$ if you place it 2.5-5cm from the Geiger, and $Y' = 2$ if you place it 0-2.5cm from the Geiger – will also be such that its actual value neutralizes the threat of $D = 1$ bringing about $E = 1$, so that **AC(M-R)** is not satisfied in the original model. The exception to this would be if the original model included a variable that represents a gerrymandered states of affairs – e.g. Y'' , which takes value $Y'' = 1$ if you place your Th-228 near the Geiger *or* Obama is US president, and $Y'' = 0$ otherwise – in which case the original model will violate (Naturalness).

7. CONCLUSION

Drawing upon PCMs, an account of PAC has been given that gives a correct treatment of probabilistic preemption on intuitive grounds. Traditional accounts of PAC misdiagnose this central test case (Menzies, 1989, 1996; Hitchcock 2004). Examination of whether PCMs can help tackle some of the other outstanding problems of PAC is warranted.

Probabilistic Actual Causation

REFERENCES

- Anscombe, E. (1971). *Causality and Determination*. Cambridge: CUP.
- Blanchard, T. and J. Schaffer (2016). Cause without Default. In H. Beebe, C. Hitchcock, and H. Price (Eds.), *Making a Difference*. Oxford: OUP.
- Goldszmidt, M. and J. Pearl (1992). Rank-Based Systems. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, San Mateo, CA, pp. 661–672. Morgan Kaufmann.
- Halpern, J. Y. (2014). Appropriate Causal Models and Stability of Causation. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*, Palo Alto, CA, pp. 198–207. AAAI Press.
- Halpern, J. Y. and C. Hitchcock (2010). Actual Causation and the Art of Modeling. In R. Dechter, H. Geffner, and J. Y. Halpern (Eds.), *Heuristics, Probability and Causality*, pp. 383–406. London: College Publications.
- Halpern, J. Y. and J. Pearl (2001). Causes and Explanations: A Structural-Model Approach. Part I: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, pp. 194–202. Morgan Kaufmann.
- Halpern, J. Y. and J. Pearl (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science* 56, 843–87.
- Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy* 98, 194–202.
- Hitchcock, C. (2004). Do All and Only Causes Raise the Probabilities of Effects? In J. Collins, N. Hall, and L. Paul (Eds.), *Causation and Counterfactuals*, pp. 403–417. Cambridge, MA: MIT Press.
- Hitchcock, C. (2007). Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review* 116, 495–532.

Luke Fenton-Glynn

- Ismael, J. (2009). Probability in Deterministic Physics. *Journal of Philosophy* 106, 89–108.
- Lewis, D. (1986). *Philosophical Papers*, Volume 2. Oxford: Oxford University Press.
- Loewer, B. (2001). Determinism and Chance. *Studies in History and Philosophy of Science Part B* 32, 609–620.
- Menzies, P. (1989). Probabilistic Causation and Causal Processes: A Critique of Lewis. *Philosophy of Science* 56, 642–663.
- Menzies, P. (1996). Probabilistic Causation and the Pre-emption Problem. *Mind* 105, 85–117.
- Paul, L. (2000). Aspect Causation. *Journal of Philosophy* 97, 235–256.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (Second ed.). Cambridge: CUP.
- Reichenbach, H. (1971). *The Direction of Time*. Mineola, NY: Dover.
- Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.
- Suppes, P. (1970). *A Probabilistic Theory of Causality*, *Acta Philosophica Fennica*. Amsterdam: North-Holland.
- Weslake, B. (2016). A Partial Theory of Actual Causation. Forthcoming in *British Journal for the Philosophy of Science*.
- Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford: OUP.

When Journal Editors Play Favorites*

Remco Heesen[†]

June 28, 2016

Abstract

Should editors of scientific journals practice triple-blind reviewing? I consider two arguments in favor of this claim. The first says that insofar as editors' decisions are affected by information they would not have had under triple-blind review, an injustice is committed against certain authors. I show that even well-meaning editors would commit this wrong and I endorse this argument.

The second argument says that insofar as editors' decisions are affected by information they would not have had under triple-blind review, it will negatively affect the quality of published papers. I distinguish between two kinds of biases that an editor might have. I show that one of them has a positive effect on quality and the other a negative one, and that the combined effect could be either positive or negative. Thus I do not endorse the second argument in general. However, I do endorse this argument for certain fields, for which I argue that the positive effect does not apply.

*Thanks to Kevin Zollman and Liam Bright for valuable comments and discussion. This work was partially supported by the National Science Foundation under grant SES 1254291.

[†]Department of Philosophy, Baker Hall 161, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA. Email: rhesen@cmu.edu

1 Introduction

Journal editors occupy an important position in the scientific landscape. By making the final decision on which papers get published in their journal and which papers do not, they have a significant influence on what work is given attention and what work is ignored in their field (Crane 1967).

In this paper I investigate the following question: should the editor be informed about the identity of the author when she is deciding whether to publish a particular paper? Under a single- or double-blind reviewing procedure, the editor has access to information about the author, whereas under a triple-blind reviewing procedure she does not. So in other words the question is: should journals practice triple-blind reviewing?

Two kinds of arguments have been given in favor of triple-blind reviewing. One focuses on the treatment of the author by the editor. On this kind of argument, revealing identity information to the editor will lead the editor to (partially) base her judgment on irrelevant information (such as the gender of the author, or whether or not the editor is friends with the author). This harms the author, and is thus bad.

The second kind of argument focuses on the effect on the journal and its readers. Again, the idea is that the editor will base her judgment on identity information if given the chance to do so. But now the further claim is that as a result the journal will accept worse papers. After all, if a decision to accept or reject a paper is influenced by the editor's biases, this suggests that a departure has been made from a putative "objectively correct" decision. This harms the readers of the journal, and is thus bad.

Here I provide a philosophical discussion of the reviewing procedure to assess these arguments. I distinguish between two different ways the editor's judgment may be affected if the author's identity is revealed to her. First, the editor may treat authors she knows differently from authors she does not know. Second, the editor may treat authors differently based on their membership of some group (e.g., gender bias). My discussion focuses on the

following three claims.

My first claim is that the first kind of differential treatment the editor may display (based on whether she knows a particular author) actually benefits rather than harms the readers of the journal. This benefit is the result of a reduction in editorial uncertainty about the quality of submitted papers when she knows their authors. I construct a model to show in a formally precise way how such a benefit might arise—surprisingly, no assumption that the scientists the editor knows are somehow “better scientists” is required—and I cite empirical evidence that such a benefit indeed does arise. However, this benefit only applies in certain fields. I argue that in other fields (in particular, mathematics and the humanities) no significant reduction of uncertainty—and hence no benefit to the readers—occurs (section 2).

My second claim is that either kind of differential treatment the editor may display (based on whether she knows authors or based on bias against certain groups) harms authors. I argue that any instance of such differential treatment constitutes an epistemic injustice in the sense of Fricker (2007) against the disadvantaged author. If the editor is to be (epistemically) just, she should prevent such differential treatment, which can be done through triple-blind reviewing. So I endorse an argument of the first of the two kinds I identified above: triple-blind reviewing is preferable because not doing so harms authors (section 3).

My third claim is that whether differential treatment also harms the journal and its readers depends on a number of factors. Differential treatment by the editor based on whether she knows a particular author may benefit readers, whereas differential treatment based on bias against certain groups may harm them. Whether there is an overall benefit or harm depends on the strength of the editor’s bias, the relative sizes of the different groups, and other factors, as I illustrate using the model. As a result I do not in general endorse the second kind of argument, that triple-blind reviewing is preferable because readers of the journal are harmed otherwise. However, I do endorse

this argument for fields like mathematics and the humanities, where I claim that the benefits of differential treatment (based on uncertainty reduction) do not apply (section 4).

Note that, in considering the ethical and epistemic effects of triple-blind reviewing, a distinction is made between the effects on the author and the effects on the readers of the journal. This reflects a growing understanding that in order to study the social epistemology of science, what is good for an individual inquirer must be distinguished from what is good for the wider scientific community (Kitcher 1993, Strevens 2003, Mayo-Wilson et al. 2011).

Zollman (2009) has studied the effects of different editorial policies on the number of papers published and the selection criteria for publication, but he does not focus specifically on the editor's decisions and the uncertainty she faces. Economists have studied models in which editor decisions play an important role (Ellison 2002, Faria 2005, Besancenot et al. 2012), but they have not distinguished between papers written by scientists the editor knows and papers by scientists unknown to her, and neither have they been concerned with biases the editor may be subject to. And some other economists have done empirical work investigating the differences between papers with and without an author-editor connection (Laband and Piette 1994, Medoff 2003, Smith and Dombrowski 1998, more on this later), but they do not provide a model that can explain these differences. This paper thus fills a gap in the literature.

2 A Model of Editor Uncertainty

As I said in the introduction, journal editors have a certain measure of power in a scientific community because they decide which papers get published.¹ An editor could use this discretionary power to the benefit of her friends or

¹Different journals may have different policies, such as one in which associate editors make the final decision for papers in their (sub)field. Here, I simply define "the editor" to be whomever makes the final decision whether to publish a particular paper.

colleagues, or to promote certain subfields or methodologies over others. This phenomenon has been called *editorial favoritism*. If anecdotal evidence is to be believed, this phenomenon is widespread. Some systematic evidence of favoritism exists as well. Bailey et al. (2008a,b) find that academics believe editorial favoritism to be fairly prevalent, with a nonnegligible percentage claiming to have perceived it firsthand. Laband (1985) and Piette and Ross (1992) find that, controlling for citation impact and various other factors, papers whose author has a connection to the journal editor are allocated more journal pages than papers by authors without such a connection.²

In this paper, I refer to the phenomenon that editors are more likely to accept papers from authors they know than papers from authors they do not know as *connection bias*.

Academics tend to disapprove of this behavior (Sherrell et al. 1989, Bailey et al. 2008a,b). In both of the studies by Bailey et al., in which subjects were asked to rate the seriousness of various potentially problematic behaviors by editors and reviewers, this disapproval was shown (using a factor analysis) to be part of a general and strong disapproval of “selfish or cliquish acts” in the peer review process. Thus it appears that the reason for the disapproval of editors publishing papers by their friends and colleagues is that it shows the editor acting on private interests, rather than displaying the disinterestedness that is the norm in science (Merton 1942).

On the other hand, if connection bias was a serious worry for authors, one would expect this to be a major consideration for them in choosing where to submit their papers (i.e., submit to journals where they know the editor), but Ziobrowski and Gibler (2000) find that this is not the case.³

²Here, page allocation is used as a proxy for journal editors’ willingness to push the paper. The more obvious variable to use here would be whether or not the paper is accepted for publication. Unfortunately, there are no empirical studies which measure the influence of a relationship between the author and the editor on acceptance decisions directly. Presumably this is because information about rejected papers is usually not available in these kinds of studies.

³In particular, authors who know an editor and thus could expect to profit from con-

Moreover, despite working scientists' disapproval, there is some evidence that connection bias improves the overall quality of accepted papers (Laband and Piette 1994, Medoff 2003, Smith and Dombrowski 1998). Does that mean scientists are misguided in their disapproval?

As indicated in the introduction, I distinguish between the effects of editors' biases on the authors of scientific papers on the one hand, and the effects on the readers of scientific journals on the other hand. In this section, I use a formal model to show that these two can come apart: connection bias may negatively affect scientists as authors while positively affecting scientists as readers. Note that in this section I focus only on connection bias. Subsequent sections consider other biases.

Consider a simplified scientific community consisting of a set of scientists. Each scientist produces a paper and submits it to the community's only journal which has one editor.

Some papers are more suitable for publication than others. I assume that this suitability for publication can be measured on a single numerical scale. For convenience I call this the *quality* of the paper. However, I remain neutral on how this notion should be interpreted, e.g., as an objective measure of the epistemic value of the paper (which is perhaps an aggregate of multiple relevant criteria), or as the number of times the paper would be cited in future papers if it was published, or as the average subjective value each member of the scientific community would assign to it if they read it.⁴

nection bias would find knowing the editor and the composition of the editorial board more generally to be important factors in deciding where to submit, contrary to Ziobrowski and Gibler's evidence (these factors are ranked twelfth and sixteenth in importance in a list of sixteen factors that might influence the decision where to submit). Similarly, authors who do not know an editor would find a lack of (perceived) connection bias and the composition of the editorial board to be important factors, but these rank only seventh and twelfth in importance in Ziobrowski and Gibler's study. In a similar survey by Mackie (1998, chapter 4), twenty percent of authors indicated that knowing the editor and/or her preferences is an important consideration in deciding where to submit a paper.

⁴For more on potential difficulties with interpreting the notion of quality, see Bright (2015).

Crucially, the editor does not know the quality of the paper at the time it is submitted. The aim of this section is to show how uncertainty about quality can lead to connection bias. To make this point as starkly as possible, I assume that the editor cares only about quality, i.e., she makes an estimate of the quality of a paper and publishes those and only those papers whose quality estimate is high.

Let q_i be the quality of the paper submitted by scientist i . Since there is uncertainty about the quality, q_i is modeled as a random variable. Since some scientists are more likely to produce high quality papers than others, the mean μ_i of this random variable may be different for each scientist. I assume that quality follows a normal distribution with fixed variance: $q_i | \mu_i \sim N(\mu_i, \sigma_{qu}^2)$.

The assumptions of normality and fixed variance are made primarily to keep the mathematics simple. Below I make similar assumptions on the distribution of average quality in the scientific community and the distribution of reviewers' estimates of the quality of a paper. I see no reason to expect the results I present below to be different when any of these assumptions are changed.

If the editor knows scientist i , she has some prior information on the average quality of scientist i 's work. This is reflected in the model by assuming that the editor knows the value of μ_i . For scientists she does not know, the editor is uncertain about the average quality of their work. All she knows is the distribution of average quality in the larger scientific community, which I also assume to be normal: $\mu_i \sim N(\mu, \sigma_{sc}^2)$.

Note that I assume the scientific community to be homogeneous: the scientific community is split in two groups (those known by the editor and those not known by the editor) but average paper quality follows the same distribution in both groups. If I assumed instead that scientists known by the editor write better papers on average the results would be qualitatively similar to those I present below. If scientists known by the editor write worse

papers on average this would affect my results. However, since most journal editors are relatively central figures in their field (Crane 1967), this would be an implausible assumption except perhaps in isolated cases.

The editor's prior beliefs about the quality of a paper submitted by some scientist i reflects this difference in information. If she knows the scientist she knows the value of μ_i , and so her prior is $\pi(q_i | \mu_i) \sim N(\mu_i, \sigma_{qu}^2)$. If the editor does not know scientist i she only knows the distribution of μ_i , rather than its exact value. Integrating out the uncertainty over μ_i yields a prior $\pi(q_i) \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2)$ for the quality of scientist i 's paper.

When the editor receives a paper she sends it out for review. In the context of this model, the main purpose of the reviewer's report is to provide an estimate of the quality of the paper. But, I assume, even after reading the paper its quality cannot be established with certainty. Thus the reviewer's estimate r_i of the quality q_i is again a random variable. I assume that the reviewer's report is unbiased, i.e., its mean is the actual quality q_i of the paper. Once again I use a normal distribution to reflect the uncertainty: $r_i | q_i \sim N(q_i, \sigma_{rv}^2)$.⁵

The editor uses the information from the reviewer's report to update her beliefs about the quality of scientist i 's paper. I assume that she does this by Bayes conditioning. Thus, her posterior beliefs about the quality of the paper are $\pi(q_i | r_i)$ if she does not know the author, and $\pi(q_i | r_i, \mu_i)$ if she does.

The posterior distributions are themselves normal distributions whose

⁵The reviewer's report could reflect the opinion of a single reviewer, or the averaged opinion of multiple reviewers. The editor could even act as a reviewer herself, in which case the report reflects her findings which she has to incorporate in her overall beliefs about the quality of the paper. The assumption I make in the text can be used to cover any of these scenarios, as long as a given journal is fairly consistent in the number of reviewers used. If the number of reviewers is frequently different for different papers (and in particular when this difference correlates with the existence or absence of a connection between editor and author) the assumption of a fixed variance in the reviewer's report is unrealistic because a report from multiple reviewers may be thought to give more accurate information (reducing the variance) than a report from a single reviewer.

mean is a weighted average of r_i and the prior mean, as given in proposition 1 (for a proof, see DeGroot 2004, section 9.5, or any other textbook that covers Bayesian statistics).

Proposition 1.

$$\pi(q_i | r_i) \sim N\left(\mu_i^U, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),$$

$$\pi(q_i | r_i, \mu_i) \sim N\left(\mu_i^K, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right),$$

where

$$\mu_i^U = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} \mu,$$

$$\mu_i^K = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} \mu_i.$$

When does the editor choose to publish a paper? Here I assume that she publishes any paper whose posterior mean is above some threshold q^* . So a paper written by a scientist unknown to the editor is published if $\mu_i^U > q^*$ and a paper written by a scientist known to the editor is published if $\mu_i^K > q^*$. This corresponds to being at least 50% confident that the paper's quality is above the threshold. Other standards could be used (risk-averse standards might require more than 50% confidence that the paper is above some threshold, while risk-loving standards might require less; in these cases the threshold value needs to be adapted to keep the total number of accepted papers constant) but for my purposes here it does not much matter.

Now compare the probability that the paper of an arbitrary scientist i unknown to the editor is published to the probability that the paper of an arbitrary scientist known by the editor is published. For this purpose it is useful to determine the probability distribution of the posterior means (see appendix A for proofs of this and subsequent results).

Proposition 2. *The posterior means are normally distributed, with $\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$. Here,*

$$\sigma_U^2 = \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} \quad \text{and} \quad \sigma_K^2 = \frac{\sigma_{qu}^4 + \sigma_{sc}^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{\sigma_{qu}^2 + \sigma_{rv}^2}.$$

Moreover, if $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$, then $\sigma_U^2 < \sigma_K^2$.

The main result of this section, which establishes the existence of connection bias in the model, is a consequence of proposition 2. It says that the editor is more likely to publish a paper written by an arbitrary author she knows than a paper written by an arbitrary author she does not know, whenever $q^* > \mu$ (for any positive value of σ_{sc}^2 and σ_{rv}^2). Since $q^* = \mu$ would mean that exactly half of all papers gets published, the condition amounts to a requirement that the journal's acceptance rate is less than 50%. This is true of most reputable journals in most fields (physics being a notable exception). When acceptance rates are above 50% editorial favoritism is also much less of a concern in the first place.

Theorem 3. *If $q^* > \mu$, $\sigma_{sc}^2 > 0$, and $\sigma_{rv}^2 > 0$, the acceptance probability for authors known to the editor is higher than the acceptance probability for authors unknown to the editor, i.e., $\Pr(\mu_i^K > q^*) > \Pr(\mu_i^U > q^*)$.*

Theorem 3 shows that in the model I presented, any journal with an acceptance rate lower than 50% will be seen to display connection bias. Thus I have established the surprising result that an editor who cares only about the quality of the papers she publishes may end up publishing more papers by her friends and colleagues than by scientists unknown to her, even if her friends and colleagues are not, as a group, better scientists than average.

Why does this surprising result hold? The theorem follows immediately from proposition 2, which says that the distribution of μ_i^U is less “spread out” than the distribution of μ_i^K ($\sigma_U^2 < \sigma_K^2$). This happens because μ_i^U is a

weighted average of μ and r_i , keeping it relatively close to the overall mean μ compared to μ_i^K , which is a weighted average of μ_i and r_i (which tend to differ from μ in the same direction).

Because the editor treats papers by authors she knows differently from papers by authors she does not know, authors unknown to the editor are arguably harmed. I pick up this point in section 3 and argue that this constitutes an epistemic injustice against those authors.

What I have shown so far is that an editor who uses information about the average quality of papers produced by scientists she knows in her acceptance decisions will find that scientists she knows produce on average more papers that meet her quality threshold. This is a subjective statement: the editor believes that more papers by scientists she knows meet her threshold. Does this translate into an objective effect? That is, does the extra information the editor has available about scientists she knows allow her to publish better papers from them than from scientists she does not know?

In order to answer this question I need to compare the average quality of accepted papers. More formally, I want to compare the expected value of the quality of a paper, conditional on meeting the publication threshold, given that the author is either known to the editor or not.

Proposition 4. *If $\sigma_{sc}^2 > 0$, and $\sigma_{rv}^2 > 0$, the average quality of accepted papers from authors known to the editor is higher than the average quality of accepted papers from authors unknown to the editor, i.e., $\mathbb{E}[q_i | \mu_i^K > q^*] > \mathbb{E}[q_i | \mu_i^U > q^*]$.*

Proposition 4 shows that the editor can use the extra information she has about scientists she knows to improve the average quality of the papers published in her journal. In other words, the surprising result is that the editor's connection bias actually benefits rather than harms the readers of the journal. It is thus fair to say that, in the model, the editor can use her connections to "identify and capture high-quality papers", as Laband and

Piette (1994) suggest.⁶

To what extent does this show that the connection bias observed in reality is the result of editors capturing high-quality papers, as opposed to editors using their position of power to help their friends? At this point the model is seen to yield an empirical prediction. If connection bias is (primarily) due to capturing high-quality papers, the quality of papers by authors the editor knows should be higher than average, as shown in the model. If, on the other hand, connection bias is (primarily) a result of the editor accepting for publication papers written by authors she knows even though they do not meet the quality standards of the journal, then the quality of papers by authors the editor knows should (presumably) be lower than average.

If subsequent citations are a good indication of the quality of a paper,⁷ a simple regression can test whether accepted papers written by authors with an author-editor connection have a higher or a lower average quality than papers without such a connection. This empirical test has been carried out a number of times, and the results univocally favor the hypothesis that editors use their connections to improve the quality of published papers (Laband and Piette 1994, Smith and Dombrowski 1998, Medoff 2003).

Note that in the above results, nothing depends on the sizes of the variances σ_{qu}^2 , σ_{sc}^2 , and σ_{rv}^2 . This is because these results are qualitative. The variances do matter when the acceptance rate and average quality of papers are compared quantitatively. For example, reducing σ_{rv}^2 (making the reviewer's report more accurate) makes the differences in the acceptance rate and average quality of papers smaller.

⁶This result applies to connection bias only. Below I consider other biases the editor might have, which yields more nuanced conclusions.

⁷Recall that I have remained neutral on how the notion of quality should be interpreted. If quality is simply defined as “the number of citations this paper would get if it were published” the connection between quality and citations is obvious. Even on other interpretations of quality, citations have frequently been viewed as a good proxy measure (Cole and Cole 1967, 1968, Medoff 2003). This practice has been defended by Cole and Cole (1971) and Clark (1957, chapter 3), and criticized by Lindsey (1989) and Heesen (forthcoming).

Note also that the results depend on the assumption that σ_{sc}^2 and σ_{rv}^2 are positive. What is the significance of these assumptions?

If $\sigma_{rv}^2 = 0$, i.e., if there is no variance in the reviewer's report, the reviewer's report describes the quality of the paper with perfect accuracy. In this case the "extra information" the editor has about authors she knows is not needed, and so there is no difference in acceptance rate or average quality based on whether the editor knows the author. But it seems unrealistic to expect reviewer's reports to be this accurate.

If $\sigma_{sc}^2 = 0$ there is either no difference in the average quality of papers produced by different authors, or learning the identity of the author does not tell the editor anything about the expected quality of that scientist's work. In this case there is no value to the editor (with regard to determining the quality of the submitted paper) in learning the identity of the author. So here also there is no difference in acceptance rate or average quality based on whether the editor knows the author.

Under what circumstances should the identity of the author be expected to tell the editor something useful about the quality of a submitted paper? This seems to be most obviously the case in the lab sciences. The identity of the author, and hence the lab at which the experiments were performed, can increase or decrease the editor's confidence that the experiments were performed correctly, including all the little checks and details that are impossible to describe in such a paper. In a scientific paper, "[a]s long as the conclusions depend at least in part on the results of some experiment, the reader must rely on the author's (and perhaps referee's) testimony that the author really performed the experiment exactly as claimed, and that it worked out as reported" (Easwaran 2009, p. 359).

But in other fields, in particular mathematics and some or all of the humanities, there is no need to rely on the author's reputation. This is because in these fields the paper itself is the contribution, so it is possible to judge papers in isolation of how or by whom they were created. Easwaran

(2009) discusses this in detail for mathematics, and briefly (in his section 4) for philosophy. And in fact there exists a norm that this is how they should be judged: “Papers will rely only on premises that the competent reader can be assumed to antecedently believe, and only make inferences that the competent reader would be expected to accept on her own consideration.” (Easwaran 2009, p. 354).

Arguably then, the advantage (see theorem 3 and proposition 4) conferred by revealing identity information about the author to the editor applies only in certain fields. The relevant fields are those where part of the information in the paper is conferred on the authority of testimony, in particular those where experimental results are reported. Even in those fields, of course, what is being testified is supposed to be reproducible by the reader. But this is still different from the case in mathematics and the humanities, where a careful reading of a paper itself constitutes a reproduction of its argument. In these latter fields there is no relevant information to be learned from the identity of the author (i.e., $\sigma_{sc}^2 = 0$), or, at least, the publishing norms in these fields suggest that their members believe this to be the case.

3 Bias As an Epistemic Injustice

The previous section discussed a formal model of editorial uncertainty about paper quality. The first main result, theorem 3, established the existence of connection bias in this model: authors known by the editor are more likely to see their paper accepted than authors unknown to the editor. The second main result, proposition 4, showed that connection bias benefits the readers of the journal by improving the average quality of accepted papers.

Despite the benefit to the readers, I claim that authors are harmed by connection bias. In this section I argue that an instance of connection bias constitutes an *epistemic injustice* in the sense of Fricker (2007). Then I argue that the editor is likely to display other biases as well, and that instances of

these also constitute epistemic injustices.

The type of epistemic justice that is relevant here is *testimonial injustice*. Fricker (2007, pp. 17–23) defines a testimonial injustice as a case where a speaker suffers a credibility deficit for which the hearer is ethically and epistemically culpable, rather than being due to innocent error.

Testimonial injustices may arise in various ways. Fricker is particularly interested in what she calls “the central case of testimonial injustice” (Fricker 2007, p. 28). This kind of injustice results from a *negative identity-prejudicial stereotype*, which is defined as follows:

A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counter-evidence owing to an ethically bad affective investment. (Fricker 2007, p. 35)

Because the stereotype is widely held, it produces *systematic* testimonial injustice: the relevant social group will suffer a credibility deficit in many different social spheres.

Applying this to the phenomenon of connection bias, it is clear that this is not an instance of the central case of testimonial injustice. This would entail that there is some negative stereotype associated with scientists unknown to the editor, as a group, which is not normally the case. So I set the central case aside (I return to it below) and focus on the question whether connection bias can produce (non-central cases of) testimonial injustice.

Suppose scientist i and scientist i' tend to produce papers of the same quality, which is above average in the population ($\mu_i = \mu_{i'} > \mu$). Suppose further that the actual papers they have produced on this occasion are of the same quality ($q_i = q_{i'}$) and have received similar reviewer reports ($r_i = r_{i'}$). If scientist i is not known to the editor, but scientist i' is, then the paper

written by scientist i' is likely to be evaluated more highly by the editor.⁸ If the publication threshold q^* is somewhere in between the two evaluations then only scientist i' will have her paper accepted.

In this example, the scientists produced papers of equal quality that were evaluated differently. So scientist i suffers a credibility deficit. This deficit is not due to innocent error, as it would be if, e.g., random variation led to different reviewer reports (i.e., $r_i < r_{i'}$). The deficit is also not due to the editor's use of generally reliable information about the two scientists, as it would be if there was a genuine difference in the average quality of the papers they produce (i.e., $\mu_i < \mu_{i'}$).

Is this credibility deficit suffered by scientist i ethically and epistemically culpable on the part of the editor? On the one hand, as I stressed in section 2, the editor is simply making maximal use of the information available to her. It just so happens that she has more information about scientists she knows than about others. But that is hardly the editor's fault: she cannot be expected to know everyone's work. Is it incumbent upon her to get to know the work of every scientist who submits a paper?

This may well be too much to ask. But an alternative option is to remove all information about the authors of submitted papers. This can be done by using a triple-blind reviewing procedure, in which the editor does not know the identity of the author, and hence is prevented from using information about scientists she knows in her evaluation. Using such a procedure, at least all scientists are treated equally: any scientist who writes a paper of a given quality has the same chance of seeing that paper accepted.

So a credibility deficit occurs which harms scientist i : her paper is rejected. Moreover, it harms her specifically as an epistemic agent: the rejection of the paper reflects a judgment of the quality of her scientific work. And

⁸The editor's posterior mean for the quality of scientist i 's paper is μ_i^U and her posterior mean for scientist i' 's paper is $\mu_{i'}^K = \mu_i^K$, with $\mu_i^U < \mu_{i'}^K$ whenever $\sigma_{sc}^2(r_i - \mu_i) < (\sigma_{qu}^2 + \sigma_{rv}^2)(\mu_i - \mu)$. The claim in the text is then justified by the fact that $\Pr(\sigma_{sc}^2(r_i - \mu_i) < (\sigma_{qu}^2 + \sigma_{rv}^2)(\mu_i - \mu) \mid \mu_i > \mu) > 1/2$, assuming $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$.

this harm could have been prevented by the editor by using a triple-blind reviewing procedure.

I conclude that the editor is ethically and epistemically culpable for this credibility deficit, and hence a testimonial injustice is committed against scientist i . However, one may insist that it cannot be the case that the editor is committing a wrong simply in virtue of using relevant information that is available to her. An evidentialist in particular may say that it cannot possibly be an epistemic wrong to take into account all relevant information.

I disagree, for the reasons just given, but I need not insist on this point. Even if it is granted that the editor does not commit an injustice by using the information that is available to her, the end result is still that scientist i is harmed as an epistemic agent. She has produced a paper of equal quality to scientist i' 's, and yet it is not published.

Moreover, the presence of scientist i' is irrelevant. Any time a paper from an author unknown to the editor is rejected which would have been accepted had the editor known the author (all else being equal), that author is harmed. So even if one insists that differential editorial treatment resulting from connection bias is not culpable on the part of the editor, connection bias still harms authors whenever it influences acceptance decisions.

In the model of section 2, and the above discussion, I assumed that connection bias is the only bias journal editors display. The literature on implicit bias suggests that this is not true. For example, “[i]f submissions are not anonymous to the editor, then the evidence suggests that women’s work will probably be judged more negatively than men’s work of the same quality” (Saul 2013, p. 45). Evidence for this claim is given by Wennerås and Wold (1997), Valian (1999, chapter 11), Steinpreis et al. (1999), Budden et al. (2008), and Moss-Racusin et al. (2012).⁹ So women scientists are at

⁹These citations show that the work of women in academia is undervalued in various ways. None of them focus specifically on editor evaluations, but they support Saul’s claim unless it is assumed that journal editors as a group are significantly less biased than other academics.

a disadvantage simply because of their gender identity. Similar biases exist based on other irrelevant aspects of scientists' identity, such as race or sexual orientation (see Lee et al. 2013, for a critical survey of various biases in the peer review system). As Crandall (1982, p. 208) puts it: "The editorial process has tended to be run as an informal, old-boy network which has excluded minorities, women, younger researchers, and those from lower-prestige institutions".

I use *identity bias* to refer to these kinds of biases. Any time a paper is rejected because of identity bias (i.e., the paper would have been accepted if the relevant part of the author's identity had been different, all else being equal), a testimonial injustice occurs for the same reasons outlined above. Moreover, here the editor is culpable for having these biases.

Unlike instances resulting from connection bias, testimonial injustices resulting from identity bias can be instances of the central case of testimonial injustice, in which the credibility deficit results from a negative identity-prejudicial stereotype. The evidence suggests that negative identity-prejudicial stereotypes affect the way people (not just men) judge women's work, even when the person judging does not consciously believe in these stereotypes. Moreover, those who think highly of their ability to judge work objectively and/or are primed with objectivity are affected more rather than less (Uhlmann and Cohen 2007, Stewart and Payne 2008, p. 1333). Similar claims plausibly hold for biases based on race or sexual orientation. Biases based on academic affiliation are not usually due to negative identity-prejudicial stereotypes, as these do not generally affect other aspects of the scientist's life.

So both connection bias and identity bias are responsible for injustices against authors. This is one way to spell out the claim that authors are harmed when journal editors do not use a triple-blind reviewing procedure. This constitutes the first kind of argument for triple-blind reviewing which I mentioned in the introduction, and which I endorse based on these consid-

erations.

4 The Effect of Bias on Quality

The second kind of argument I mentioned in the introduction claims that failing to use triple-blind reviewing harms the journal and its readers, because it would lower the average quality of accepted papers. In section 2 I argued that connection bias actually has the opposite effect: it increases average quality. In this section I complicate the model to include identity bias.

Recall that the editor displays identity bias if she is more or less likely to publish papers from a certain group of scientists based on some aspect of their identity, e.g., their gender. I incorporate this in the model by assuming the editor consistently undervalues members of one group (and overvalues the others). More precisely, she believes the average quality of papers produced by any scientist i from the group she is biased against to be lower than it really is by some constant quantity ε . Conversely, the average quality of papers written by any scientist not belonging to this group is raised by δ .¹⁰ So the editor has a different prior for the two groups; I use π_A to denote her prior for the quality of papers written by scientists she is biased against, and π_F for her prior for scientists she is biased in favor of.

As before, the editor may be familiar with a given scientist's work (i.e., she knows the average quality of that scientist's papers) or not. So there are now four groups. If scientist i is known to the editor and belongs to the stigmatized group the editor's prior distribution on the quality of scientist i 's paper is $\pi_A(q_i | \mu_i) \sim N(\mu_i - \varepsilon, \sigma_{qu}^2)$. If scientist i is known to the editor but is not in the stigmatized group the prior is $\pi_F(q_i | \mu_i) \sim N(\mu_i + \delta, \sigma_{qu}^2)$. If

¹⁰This is a simplifying assumption: one could imagine having biases against multiple groups of different strengths, or biases whose strength has some random variation, or biases which intersect in various ways (Collins and Chepp 2013, Bright et al. 2016). However, the assumption in the main text suffices to make the point I want to make. It should be fairly straightforward to extend my results to more complicated cases like the ones just described.

scientist i is not known to the editor and is in the stigmatized group the prior is $\pi_A(q_i) \sim N(\mu - \varepsilon, \sigma_{qu}^2 + \sigma_{sc}^2)$. And if scientist i is not known to the editor and not in the stigmatized group the prior is $\pi_F(q_i) \sim N(\mu + \delta, \sigma_{qu}^2 + \sigma_{sc}^2)$.¹¹

The next few steps in the development are analogous to that in section 2. After the reviewer's report comes in the editor updates her beliefs about the quality of the paper, yielding the following posterior distributions.

Proposition 5.

$$\begin{aligned}\pi_A(q_i | r_i, \mu_i) &\sim N\left(\mu_i^{KA}, \frac{\sigma_{qu}^2 \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right), \\ \pi_F(q_i | r_i, \mu_i) &\sim N\left(\mu_i^{KF}, \frac{\sigma_{qu}^2 \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right), \\ \pi_A(q_i | r_i) &\sim N\left(\mu_i^{UA}, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2) \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right), \\ \pi_F(q_i | r_i) &\sim N\left(\mu_i^{UF}, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2) \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),\end{aligned}$$

where

$$\begin{aligned}\mu_i^{KA} &= \mu_i^K - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, & \mu_i^{KF} &= \mu_i^K + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \\ \mu_i^{UA} &= \mu_i^U - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, & \mu_i^{UF} &= \mu_i^U + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}.\end{aligned}$$

As before, the paper is published if the posterior mean (μ_i^{KA} , μ_i^{KF} , μ_i^{UA} , or μ_i^{UF}) exceeds the threshold q^* . The respective distributions of the posterior

¹¹Note that I assume that the editor displays bias against scientists in the stigmatized group regardless of whether she knows them or not. Under a reviewing procedure that is not triple-blind, the editor learns at least the name and affiliation of any scientist who submits a paper. This information is usually sufficient to determine with reasonable certainty the scientist's gender. So at least for gender bias it seems reasonable to expect the editor to display bias even against scientists she does not know. Conversely, because negative identity-prejudicial stereotypes can work unconsciously, it does not seem reasonable to expect that the editor can withhold her bias from scientists she knows.

means determine how likely this is. These distributions are given in the next proposition.

Proposition 6. *The posterior means are normally distributed, with*

$$\begin{aligned}\mu_i^{KA} &\sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right), \\ \mu_i^{KF} &\sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right), \\ \mu_i^{UA} &\sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right), \\ \mu_i^{UF} &\sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right).\end{aligned}$$

This yields the within-group acceptance rates and the unsurprising result that the editor is less likely to publish papers by scientists she is biased against.

Theorem 7. *If $\varepsilon > 0$, $\delta > 0$, and $\sigma_{rv}^2 > 0$, the acceptance probability for authors the editor is biased against is lower than the acceptance probability for authors the editor is biased in favor of (keeping fixed whether or not the editor knows the author). That is,*

$$\Pr(\mu_i^{KA} > q^*) < \Pr(\mu_i^{KF} > q^*) \quad \text{and} \quad \Pr(\mu_i^{UA} > q^*) < \Pr(\mu_i^{UF} > q^*).$$

Theorem 7 establishes the existence of identity bias in the model: authors that are subject to a negative identity-prejudicial stereotype are less likely to see their paper accepted than authors who are not. As I argued in section 3, whenever a paper is rejected due to identity bias this constitutes a testimonial injustice against the author.

Now I turn my attention to the effect that identity bias has on the average quality of accepted papers. In the current version of the model there is both

connection bias and identity bias. Connection bias has been shown to have a positive effect on average quality (see section 2). Whether the net effect of connection bias and identity bias is positive or negative depends on various parameters, as I illustrate below.

The benchmark for judging the average quality of accepted papers under a procedure subject to connection bias and identity bias is a *triple-blind reviewing procedure* under which the editor is not informed of the identity of the scientist. As a result, she is both unable to use information about the average quality of a given scientist's papers and unable to display bias against scientists based on their identity.

Under this triple-blind procedure, the editor's prior distribution for the quality of any submitted paper is $\pi(q_i) \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2)$, i.e., the prior I used in section 2 when the author was unknown to the editor. Hence, under this procedure, the posterior is $\pi(q_i | r_i)$, the posterior mean is $\mu_i^U \sim N(\mu, \sigma_U^2)$, the probability of acceptance is $\Pr(\mu_i^U > q^*)$ and the average quality of accepted papers is $\mathbb{E}[q_i | \mu_i^U > q^*]$.

In contrast, I refer to the reviewing procedure that is subject to connection bias and identity bias as the *non-blind procedure*. The overall probability that a paper is accepted under the non-blind procedure depends on the relative sizes of the four groups. I use p_{KA} to denote the fraction of scientists known to the editor that she is biased against, p_{KF} for the fraction known to the editor that she is biased in favor of, p_{UA} for unknown scientists biased against, and p_{UF} for unknown scientists biased in favor of. These fractions are nonnegative and sum to one.

Let A_i denote the event that scientist i 's paper is accepted under the non-blind procedure. The overall probability of acceptance under this procedure is

$$\begin{aligned} \Pr(A_i) &= p_{KA} \Pr(\mu_i^{KA} > q^*) + p_{KF} \Pr(\mu_i^{KF} > q^*) \\ &\quad + p_{UA} \Pr(\mu_i^{UA} > q^*) + p_{UF} \Pr(\mu_i^{UF} > q^*). \end{aligned}$$

The average quality of accepted papers can then be written as $\mathbb{E}[q_i | A_i]$. I want to compare $\mathbb{E}[q_i | A_i]$ to $\mathbb{E}[q_i | \mu_i^U > q^*]$, the average quality of accepted papers under a triple-blind procedure.¹²

In the remainder of this section I assume that the editor's biases are such that she believes the average quality of all submitted papers to be equal to μ . In other words, her bias against the stigmatized group is canceled out on average by her bias in favor of those not in the stigmatized group, weighted by the relative sizes of those groups:

$$(p_{KA} + p_{UA}) \varepsilon = (p_{KF} + p_{UF}) \delta.$$

I use the above equation to fix the value of δ , reducing the number of free parameters by one. The equation amounts to a kind of commensurability requirement for the two procedures because it guarantees that the editor perceives the average quality of submitted papers to be the same regardless of whether or not a triple-blind procedure is used.

As far as I can tell there are no interesting general conditions on the parameter values that determine whether the non-blind procedure or the triple-blind procedure will lead to a higher average quality of accepted papers. The question I will explore now, using some numerical examples, is how biased the editor needs to be for the epistemic costs of her identity bias to outweigh the epistemic benefits resulting from connection bias.

In order to generate numerical data values have to be chosen for the

¹²Expressions for $\Pr(A_i)$ and $\mathbb{E}[q_i | A_i]$ using only the parameter values and standard functions are given in lemma 11 in appendix A. These expressions are used to generate the numerical results below.

parameters. First I set $\mu = 0$ and $q^* = 2$. Since quality is an interval scale in this model, these choices are arbitrary. For the variances σ_{qu}^2 , σ_{sc}^2 , and σ_{rv}^2 , I choose a “small” and a “large” value (1 and 4 respectively).

For the sizes of the four groups, I assume that there is no correlation between whether the editor knows an author and whether the editor has a bias against that author (so, e.g., the percentage of women among scientists the editor knows is equal to the percentage of women among scientists the editor does not know). I consider two cases for the editor’s identity bias: either she is biased against half the set of authors (and so biased in favor of the other half) or the group she is biased against is a 30 % minority.¹³ Similarly, I consider the case in which the editor knows half of all scientists submitting papers, and the case in which the editor knows 30 % of them.

As a result, there are 32 possible settings of the parameters (2^3 choices for the variances times 2^2 choices for the group sizes). Whether the triple-blind procedure or the non-blind procedure is epistemically preferable depends on the value of ε (and the value of δ determined thereby).

It follows from proposition 4 that when $\varepsilon = 0$ the non-blind procedure helps rather than harms the readers of the journal by increasing average quality relative to the triple-blind procedure. If ε is positive but relatively small, this remains true, but when ε is relatively big, the non-blind procedure harms the readers. This is because the average quality of published papers under the non-blind procedure decreases continuously as ε increases (I do not prove this, but it is easily checked for the 32 cases I consider).

The interesting question, then, is where the turning point lies. How big does the editor’s bias need to be in order for the negative effects of identity bias on quality to cancel out the positive effects of connection bias?

¹³Bruner and O’Connor (forthcoming) note that certain dynamics in academic life can lead to identity bias against groups as a result of the mere fact that they are a minority. Here I consider both the case where the stigmatized group is a minority (and is possibly stigmatized as a result of being a minority, as Bruner and O’Connor suggest) and the case where it is not (and so presumably the negative identity-prejudicial stereotype has some other source).

I determine the value of ε for which the average quality of published papers under the non-blind procedure and the triple-blind procedure is the same for each of the 32 cases. But reporting these numbers directly does not seem particularly useful, as ε is measured in “quality points” which do not have a clear interpretation outside of the model.

To give a more meaningful interpretation of these values of ε as measuring “size of bias”, I calculate the average rate of acceptance of papers from authors the editor is biased against and the average rate of acceptance of papers from authors the editor is biased in favor of.¹⁴ The difference between these numbers gives an indication of the size of the editor’s bias: it measures (in percentage points, abbreviated pp) how many more papers the editor accepts from authors she is biased in favor of, compared to those she is biased against.

This difference is reported for the 32 cases in figure 1. To provide a sense of scale for these numbers, I plot them against the acceptance rate that the triple-blind procedure would have for those values of the parameters, i.e., $\Pr(\mu_i^U > q^*)$.

Already with this small sample of 32 cases, a large variation of results can be observed. I illustrate this by looking at two cases in detail.

First, suppose that $\sigma_{qu}^2 = \sigma_{sc}^2 = 1$ and $\sigma_{rv}^2 = 4$. In this extreme case the triple-blind procedure has an acceptance rate as low as 0.72%. If the groups are all of equal size ($p_{KA} = p_{KF} = p_{UA} = p_{UF} = 1/4$) then under the non-blind procedure the acceptance rate for authors the editor is biased in favor of needs to be as much as 2.66 pp higher than the acceptance rate for authors the editor is biased against, in order for the average quality under

¹⁴These are calculated without regard for whether the editor knows the author or not. In particular, the rate of acceptance for authors the editor is biased against is

$$\frac{p_{KA} \Pr(\mu_i^{KA} > q^*) + p_{UA} \Pr(\mu_i^{UA} > q^*)}{p_{KA} + p_{UA}}, \text{ and } \frac{p_{KF} \Pr(\mu_i^{KF} > q^*) + p_{UF} \Pr(\mu_i^{UF} > q^*)}{p_{KF} + p_{UF}}$$

is the rate of acceptance for authors the editor is biased in favor of.

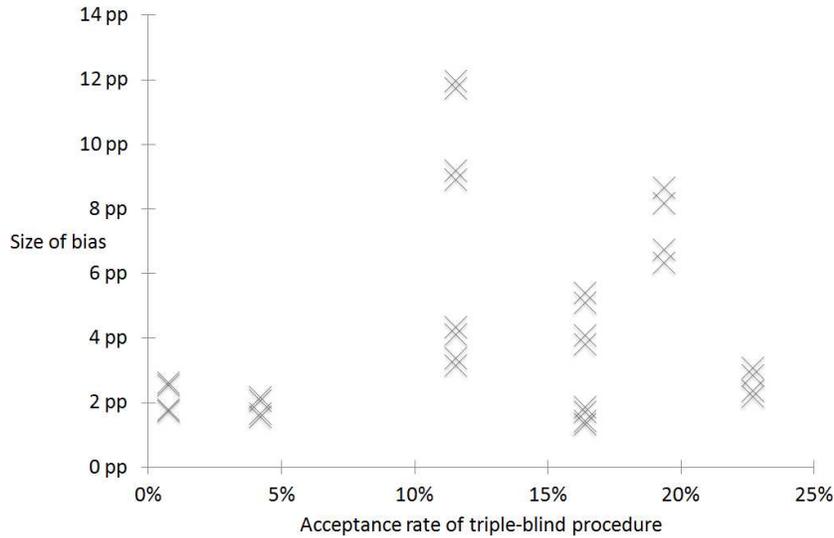


Figure 1: The minimum size of the editor’s bias such that the quality costs of the non-blind procedure outweigh its benefits (given as a percentage point difference in acceptance rates), in 32 cases, plotted as a function of the acceptance rate of the corresponding triple-blind procedure.

the two procedures to be equal. Clearly a 2.66 pp bias is very large for a journal that only accepts less than 1% of papers. If the bias is any less than that there is no harm to the readers in using the non-blind procedure.

Second, suppose that $\sigma_{qu}^2 = \sigma_{sc}^2 = 4$ and $\sigma_{rv}^2 = 1$. Then the triple-blind procedure has an acceptance rate of 22.66%. If, moreover, the editor knows relatively few authors ($p_{KA} = p_{KF} = 0.15$, $p_{UA} = p_{UF} = 0.35$) then the acceptance rate for authors the editor is biased in favor of needs to be only 2.23 pp higher than the acceptance rate for authors the editor is biased against, in order for the quality costs of the non-blind procedure to outweigh its benefits. For a journal accepting about 23% of papers that means that even if the identity bias of the editor is relatively mild the journal’s readers are harmed if the non-blind procedure is used.

Based on these results, and the fact that the parameter values are unlikely to be known in practice, it is unclear whether the non-blind procedure

or the triple-blind procedure will lead to a higher average quality of published papers for any particular journal.¹⁵ So in general it is not clear that an argument that the non-blind procedure harms the journal's readers can be made. At the same time, a general argument that the non-blind procedure helps the readers is not available either. Given this, I am inclined to recommend a triple-blind procedure for all journals because not doing so harms the authors.

If there was reason to believe that the editor's bias was very small, there might be a case for the non-blind procedure using considerations of average quality. Based on the empirical evidence I cited in section 3, it seems unlikely that any editor could make such a case convincingly today. But if identity bias were someday to be eliminated or severely mitigated, this question may be worth revisiting.

So far I have argued in this section that in the presence of the positive effect of connection bias on quality, the net effect of connection bias and identity bias on quality is unclear. But I argued in section 2 that the positive effect of connection bias may only exist in certain fields. In fields where papers rely partially on the author's testimony there is value in knowing the identity of the author. But in other fields such as mathematics and some of the humanities testimony is not taken to play a role—the paper itself constitutes the contribution to the field—and so arguably there is no value in knowing the identity of the author.

In those fields, then, there is no quality benefit from connection bias, but there is still a quality cost from identity bias. So here the strongest case for the triple-blind procedure emerges, as the non-blind procedure harms both authors and readers.

¹⁵Note that the evidence collected by Laband and Piette (1994) does not help settle this question, as they do not directly compare the triple-blind and the non-blind procedure. Their evidence supports a positive epistemic effect of connection bias, but not a verdict on the overall epistemic effect of triple-blinding.

5 Conclusion

In this paper I have considered two types of arguments for triple-blind review: one based on the consequences for the author and one based on the consequences for the readers of the journal.

I have argued that the non-blind procedure introduces differential treatment of scientific authors. In particular, editors are more likely to publish papers by authors they know (connection bias, theorem 3) and less likely to publish papers by authors they apply negative identity-prejudicial stereotypes to (identity bias, theorem 7). Whenever a paper is rejected as a result of one of these biases an epistemic injustice (in the sense of Fricker 2007) is committed against the author. This is an argument in favor of triple-blinding based on consequences for the author.

From the readers' perspective the story is more mixed. Generally speaking connection bias has a positive effect on the quality of published papers and identity bias a negative one. Thus whether the readers are better off under the triple-blind procedure depends on how exactly these effects trade off, which is highly context-dependent, or so I have argued. This yields a more nuanced view than that suggested by either Laband and Piette (1994), who focus only on connection bias, or by the argument for triple-blinding based on the consequences for the readers, which focuses only on identity bias.

However, in mathematics and some of the humanities there is arguably no positive quality effect from connection bias, as knowing about an author's other work is not taken to be relevant (Easwaran 2009). So here the negative effect of identity bias is the only relevant consideration from the readers' perspective. In this situation, considerations concerning the consequences for the author and considerations concerning the consequences for the readers point in the same direction: in favor of triple-blind review.

A The Acceptance Probability and the Average Quality of Papers

Proposition 2. $\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$. Moreover, $\sigma_U^2 < \sigma_K^2$ whenever $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$.

Proof. First consider the distribution of r_i . Since $r_i \mid q_i \sim N(q_i, \sigma_{rv}^2)$, $q_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2)$, and $\mu_i \sim N(\mu, \sigma_{sc}^2)$, it follows that $r_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2 + \sigma_{rv}^2)$ and $r_i \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)$.

The latter can be used straightforwardly to determine the distribution of μ_i^U . Since $r_i - \mu \sim N(0, \sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)$ it follows that

$$\frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}(r_i - \mu) \sim N\left(0, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right) \sim N(0, \sigma_U^2).$$

The result follows because μ is a constant and

$$\mu_i^U = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\mu = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}(r_i - \mu) + \mu.$$

Determining the distribution of μ_i^K is slightly trickier because there are two random variables involved: r_i and μ_i . As noted above, $r_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2 + \sigma_{rv}^2)$. Thus, writing $X_i = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}(r_i - \mu_i)$,

$$X_i \mid \mu_i \sim N\left(0, \frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}\right).$$

Since

$$\mu_i^K = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\mu_i = X_i + \mu_i$$

it remains to determine the convolution of X_i and μ_i . This can be done using

the moment-generating function and the law of total expectation. Recall that the moment-generating function of an $N(m, s^2)$ distribution is given by $M(t) = \exp\{mt + \frac{1}{2}s^2t^2\}$. So the moment-generating function of μ_i^K is

$$\begin{aligned}
\mathbb{E}[\exp\{t\mu_i^K\}] &= \mathbb{E}[\exp\{t(X_i + \mu_i)\}] \\
&= \mathbb{E}[\mathbb{E}[\exp\{tX_i + t\mu_i\} \mid \mu_i]] \\
&= \mathbb{E}[\exp\{t\mu_i\}\mathbb{E}[\exp\{tX_i\} \mid \mu_i]] \\
&= \exp\left\{0t + \frac{1}{2}\frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2\right\} \mathbb{E}[\exp\{t\mu_i\}] \\
&= \exp\left\{\frac{1}{2}\frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2 + \mu t + \frac{1}{2}\sigma_{sc}^2t^2\right\} \\
&= \exp\left\{\mu t + \frac{1}{2}\frac{\sigma_{qu}^4 + \sigma_{sc}^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2\right\},
\end{aligned}$$

which is exactly the moment-generating function of the desired normal distribution.

Finally, note that

$$\begin{aligned}
\sigma_U^2 &= \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{(\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)(\sigma_{qu}^2 + \sigma_{rv}^2)}, \\
\sigma_K^2 &= \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2(\sigma_{qu}^2 + \sigma_{rv}^2) + \sigma_{sc}^2\sigma_{rv}^4}{(\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)(\sigma_{qu}^2 + \sigma_{rv}^2)}.
\end{aligned}$$

So $\sigma_U^2 < \sigma_K^2$ whenever $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$ (and $\sigma_U^2 = \sigma_K^2$ otherwise, assuming the expressions are well-defined in that case). \square

Theorem 3. $\Pr(\mu_i^K > q^*) > \Pr(\mu_i^U > q^*)$ if $q^* > \mu$, $\sigma_{sc}^2 > 0$, and $\sigma_{rv}^2 > 0$.

Proof. It follows from proposition 2 that

$$\Pr(\mu_i^K > q^*) = 1 - \Phi\left(\frac{q^* - \mu}{\sigma_K}\right) \text{ and } \Pr(\mu_i^U > q^*) = 1 - \Phi\left(\frac{q^* - \mu}{\sigma_U}\right),$$

where Φ is the distribution function (or cumulative density function) of a standard normal distribution. Since Φ is (strictly) increasing in its argument, and $\sigma_K > \sigma_U$ by proposition 2, the theorem follows immediately. \square

In order to prove proposition 4 a number of intermediate results are needed.

Lemma 8.

$$\begin{aligned} \mathbb{E}[q_i \mid \mu_i^U > q^*] &= \mathbb{E}[\mu_i^U \mid \mu_i^U > q^*], \\ \mathbb{E}[q_i \mid \mu_i^K > q^*] &= \mathbb{E}[\mu_i^K \mid \mu_i^K > q^*]. \end{aligned}$$

Proof. Because μ_i^U is simply an (invertible) transformation of r_i , it follows that

$$q_i \mid \mu_i^U \sim q_i \mid r_i \sim N\left(\mu_i^U, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right).$$

The distribution of $q_i \mid \mu_i^K$ is a little trickier to find, because μ_i^K is a linear combination of two random variables, r_i and μ_i , and it is not obvious that learning μ_i^K is as informative as learning both r_i and μ_i . But using the known distributions of $q_i \mid \mu_i$ and $\mu_i^K \mid q_i, \mu_i$ and integrating out μ_i it can be shown that

$$q_i \mid \mu_i^K \sim q_i \mid r_i, \mu_i \sim N\left(\mu_i^K, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right).$$

The important point here is that $\mathbb{E}[q_i \mid \mu_i^x] = \mu_i^x$ both for $x = U$ and $x = K$.

Now the law of total expectation can be used to establish that

$$\mathbb{E}[q_i \mid \mu_i^x > q^*] = \mathbb{E}[\mathbb{E}[q_i \mid \mu_i^x] \mid \mu_i^x > q^*] = \mathbb{E}[\mu_i^x \mid \mu_i^x > q^*],$$

for $x = U, K$. □

Let $X \sim N(\mu, \sigma^2)$ be a normally distributed random variable. Then $X \mid X > a$ follows a *left-truncated normal distribution*, with left-truncation point a . As a result of lemma 8 I am interested in the mean of left-truncated normal distributions. According to, e.g., Johnson et al. (1994, chapter 13, section 10.1), this mean can be expressed as

$$\mathbb{E}[X \mid X > a] = \mu + \sigma R\left(\frac{a - \mu}{\sigma}\right). \quad (1)$$

Here

$$R(x) = \frac{\phi(x)}{1 - \Phi(x)}$$

for all $x \in \mathbb{R}$, where ϕ is the probability density function of the standard normal distribution, and Φ is its distribution function. R is the inverse of what is known in the literature (e.g., Gordon 1941) as *Mills' ratio*.

It follows from the definitions that $R(x) > 0$ for all $x \in \mathbb{R}$ and that

$$R'(x) = R(x)^2 - xR(x). \quad (2)$$

Proposition 9 (Gordon (1941)). *For all $x > 0$, $R(x) < \frac{x^2+1}{x}$.*

Proposition 9 can be used to establish the next result.

Proposition 10. *If $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\mu, s^2)$ with $s > \sigma > 0$ then $\mathbb{E}[Y \mid Y > a] > \mathbb{E}[X \mid X > a]$.*

Proof. It suffices to show that the derivative $\frac{\partial}{\partial \sigma} \mathbb{E}[X \mid X > a]$ is positive for all $\sigma > 0$. Differentiating equation (1) (using equation (2)) yields

$$\frac{\partial}{\partial \sigma} \mathbb{E}[X | X > a] = \left(\left(\frac{a - \mu}{\sigma} \right)^2 + 1 \right) R \left(\frac{a - \mu}{\sigma} \right) - \frac{a - \mu}{\sigma} R \left(\frac{a - \mu}{\sigma} \right)^2.$$

Since $R \left(\frac{a - \mu}{\sigma} \right) > 0$, $\frac{\partial}{\partial \sigma} \mathbb{E}[X | X > a] > 0$ if and only if

$$\left(\frac{a - \mu}{\sigma} \right)^2 + 1 - \frac{a - \mu}{\sigma} R \left(\frac{a - \mu}{\sigma} \right) > 0.$$

This is true whenever $\frac{a - \mu}{\sigma} \leq 0$ because then both terms in the sum are positive. Proposition 9 guarantees that it is true whenever $\frac{a - \mu}{\sigma} > 0$ as well. \square

Proposition 4. $\mathbb{E}[q_i | \mu_i^K > q^*] > \mathbb{E}[q_i | \mu_i^U > q^*]$ whenever $\sigma_{sc}^2 > 0$, and $\sigma_{rv}^2 > 0$.

Proof. By lemma 8,

$$\begin{aligned} \mathbb{E}[q_i | \mu_i^U > q^*] &= \mathbb{E}[\mu_i^U | \mu_i^U > q^*], \\ \mathbb{E}[q_i | \mu_i^K > q^*] &= \mathbb{E}[\mu_i^K | \mu_i^K > q^*]. \end{aligned}$$

By proposition 2, $\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$, with $\sigma_U < \sigma_K$. Hence the conditions of proposition 10 are satisfied, and the result follows. \square

Proposition 6.

$$\begin{aligned} \mu_i^{KA} &\sim N \left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2 \right), \\ \mu_i^{KF} &\sim N \left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2 \right), \\ \mu_i^{UA} &\sim N \left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2 \right), \\ \mu_i^{UF} &\sim N \left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2 \right). \end{aligned}$$

Proof. Since μ_i^{KA} and μ_i^{KF} are simply μ_i^K shifted by a constant (see proposition 5) they follow the same distribution as μ_i^K except that its mean is shifted by the same constant. Similarly μ_i^{UA} and μ_i^{UF} are just μ_i^U shifted by a constant. So the results follow from proposition 2. \square

For notational convenience, I introduce q^{KA} , q^{KF} , q^{UA} , and q^{UF} , defined by

$$\begin{aligned} q^{KA} &= q^* + \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, & q^{KF} &= q^* - \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \\ q^{UA} &= q^* + \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, & q^{UF} &= q^* - \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}. \end{aligned}$$

Theorem 7. *If $\varepsilon > 0$, $\delta > 0$, and $\sigma_{rv}^2 > 0$,*

$$\Pr(\mu_i^{KA} > q^*) < \Pr(\mu_i^{KF} > q^*) \text{ and } \Pr(\mu_i^{UA} > q^*) < \Pr(\mu_i^{UF} > q^*).$$

Proof. For the first inequality, note that

$$\Pr(\mu_i^{KA} > q^*) = 1 - \Phi\left(\frac{q^{KA} - \mu}{\sigma_K}\right) < 1 - \Phi\left(\frac{q^{KF} - \mu}{\sigma_K}\right) = \Pr(\mu_i^{KF} > q^*).$$

The equalities follow from the distributions of the posterior means established in proposition 6. The inequality follows from the fact that Φ is strictly increasing in its argument. By the same reasoning,

$$\Pr(\mu_i^{UA} > q^*) = 1 - \Phi\left(\frac{q^{UA} - \mu}{\sigma_U}\right) < 1 - \Phi\left(\frac{q^{UF} - \mu}{\sigma_U}\right) = \Pr(\mu_i^{UF} > q^*). \quad \square$$

Lemma 11.

$$\begin{aligned} \Pr(A_i) &= p_{KA} \left(1 - \Phi \left(\frac{q^{KA} - \mu}{\sigma_K} \right) \right) + p_{KF} \left(1 - \Phi \left(\frac{q^{KF} - \mu}{\sigma_K} \right) \right) \\ &\quad + p_{UA} \left(1 - \Phi \left(\frac{q^{UA} - \mu}{\sigma_U} \right) \right) + p_{UF} \left(1 - \Phi \left(\frac{q^{UF} - \mu}{\sigma_U} \right) \right). \\ \mathbb{E}[q_i | A_i] &= \mu + \frac{\sigma_K}{\Pr(A_i)} \left(p_{KA} \phi \left(\frac{q^{KA} - \mu}{\sigma_K} \right) + p_{KF} \phi \left(\frac{q^{KF} - \mu}{\sigma_K} \right) \right) \\ &\quad + \frac{\sigma_U}{\Pr(A_i)} \left(p_{UA} \phi \left(\frac{q^{UA} - \mu}{\sigma_U} \right) + p_{UF} \phi \left(\frac{q^{UF} - \mu}{\sigma_U} \right) \right). \end{aligned}$$

Proof. The expression for $\Pr(A_i)$ follows immediately from the distributions of the posterior means established in proposition 6.

To get an expression for $\mathbb{E}[q_i | A_i]$, consider first the average quality of scientist i 's paper given that it is accepted and given that scientist i is in the group of scientists known to the editor that the editor is biased against. This average quality is

$$\begin{aligned} \mathbb{E}[q_i | \mu_i^{KA} > q^*] &= \mathbb{E}[q_i | \mu_i^K > q^{KA}] = \mathbb{E}[\mu_i^K | \mu_i^K > q^{KA}] \\ &= \mu + \sigma_K R \left(\frac{q^{KA} - \mu}{\sigma_K} \right), \end{aligned}$$

where the first equality simply rewrites the inequality $\mu_i^{KA} > q^*$ in a more convenient form, the second equality uses lemma 8, and the third equality uses equation 1. Similarly,

$$\begin{aligned}\mathbb{E}[q_i \mid \mu_i^{KF} > q^*] &= \mu + \sigma_K R \left(\frac{q^{KF} - \mu}{\sigma_K} \right), \\ \mathbb{E}[q_i \mid \mu_i^{UA} > q^*] &= \mu + \sigma_U R \left(\frac{q^{UA} - \mu}{\sigma_U} \right), \\ \mathbb{E}[q_i \mid \mu_i^{UF} > q^*] &= \mu + \sigma_U R \left(\frac{q^{UF} - \mu}{\sigma_U} \right).\end{aligned}$$

The average quality of accepted papers $\mathbb{E}[q_i \mid A_i]$ is a weighted sum of these expectations. The weights are given by the proportion of accepted papers that are written by a scientist in that particular group. For example, authors known to the editor that she is biased against form a $p_{KA} \Pr(\mu_i^{KA} > q^*) / \Pr(A_i)$ proportion of accepted papers. Hence

$$\begin{aligned}\mathbb{E}[q_i \mid A_i] &= \frac{1}{\Pr(A_i)} p_{KA} \Pr(\mu_i^{KA} > q^*) \mathbb{E}[q_i \mid \mu_i^{KA} > q^*] \\ &\quad + \frac{1}{\Pr(A_i)} p_{KF} \Pr(\mu_i^{KF} > q^*) \mathbb{E}[q_i \mid \mu_i^{KF} > q^*] \\ &\quad + \frac{1}{\Pr(A_i)} p_{UA} \Pr(\mu_i^{UA} > q^*) \mathbb{E}[q_i \mid \mu_i^{UA} > q^*] \\ &\quad + \frac{1}{\Pr(A_i)} p_{UF} \Pr(\mu_i^{UF} > q^*) \mathbb{E}[q_i \mid \mu_i^{UF} > q^*] \\ &= \mu + \frac{\sigma_K}{\Pr(A_i)} \left(p_{KA} \phi \left(\frac{q^{KA} - \mu}{\sigma_K} \right) + p_{KF} \phi \left(\frac{q^{KF} - \mu}{\sigma_K} \right) \right) \\ &\quad + \frac{\sigma_U}{\Pr(A_i)} \left(p_{UA} \phi \left(\frac{q^{UA} - \mu}{\sigma_U} \right) + p_{UF} \phi \left(\frac{q^{UF} - \mu}{\sigma_U} \right) \right). \quad \square\end{aligned}$$

References

Charles D. Bailey, Dana R. Hermanson, and Timothy J. Louwers. An examination of the peer review process in accounting journals. *Jour-*

- nal of Accounting Education*, 26(2):55–72, 2008a. ISSN 0748-5751. doi: 10.1016/j.jaccedu.2008.04.001. URL <http://www.sciencedirect.com/science/article/pii/S0748575108000201>.
- Charles D. Bailey, Dana R. Hermanson, and James G. Tompkins. The peer review process in finance journals. *Journal of Financial Education*, 34: 1–27, 2008b. ISSN 0093-3961. URL <http://www.jstor.org/stable/41948838>.
- Damien Besancenot, Kim V. Huynh, and Joao R. Faria. Search and research: the influence of editorial boards on journals' quality. *Theory and Decision*, 73(4):687–702, 2012. ISSN 0040-5833. doi: 10.1007/s11238-012-9314-7. URL <http://dx.doi.org/10.1007/s11238-012-9314-7>.
- Liam Kofi Bright. Against candidate quality. Manuscript, 2015. URL https://www.academia.edu/11673059/Against_Candidate_Quality.
- Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. Causally interpreting intersectionality theory. *Philosophy of Science*, 83(1):60–81, 2016. ISSN 00318248. URL <http://www.jstor.org/stable/10.1086/684173>.
- 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*. Oxford University Press, Oxford, forthcoming. URL <http://philpapers.org/rec/BRUPBA-2>.
- Amber E. Budden, Tom Tregenza, Lonnie W. Aarssen, Julia Koricheva, Roosa Leimu, and Christopher J. Lortie. Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, 23(1):4–6, 2008. ISSN 0169-5347. doi: 10.1016/j.tree.2007.07.008. URL <http://www.sciencedirect.com/science/article/pii/S0169534707002704>.

- Kenneth E. Clark. *America's Psychologists: A Survey of a Growing Profession*. American Psychological Association, Washington, 1957.
- Jonathan R. Cole and Stephen Cole. Measuring the quality of sociological research: Problems in the use of the "Science Citation Index". *The American Sociologist*, 6(1):23–29, 1971. ISSN 00031232. URL <http://www.jstor.org/stable/27701705>.
- Stephen Cole and Jonathan R. Cole. Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, 32(3):377–390, 1967. ISSN 00031224. URL <http://www.jstor.org/stable/2091085>.
- Stephen Cole and Jonathan R. Cole. Visibility and the structural bases of awareness of scientific research. *American Sociological Review*, 33(3):397–413, 1968. ISSN 00031224. URL <http://www.jstor.org/stable/2091914>.
- Patricia Hill Collins and Valerie Chepp. Intersectionality. In Georgina Waylen, Karen Celis, Johanna Kantola, and S. Laurel Weldon, editors, *The Oxford Handbook of Gender and Politics*, chapter 2, pages 57–87. Oxford University Press, Oxford, 2013. ISBN 0199751455.
- Rick Crandall. Editorial responsibilities in manuscript review. *Behavioral and Brain Sciences*, 5:207–208, Jun 1982. ISSN 1469-1825. doi: 10.1017/S0140525X00011316. URL http://journals.cambridge.org/article_S0140525X00011316.
- 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>.
- Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, New Jersey, 2004.

- Kenny Easwaran. Probabilistic proofs and transferability. *Philosophia Mathematica*, 17(3):341–362, 2009. doi: 10.1093/phimat/nkn032. URL <http://phimat.oxfordjournals.org/content/17/3/341.abstract>.
- Glenn Ellison. Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy*, 110(5):994–1034, 2002. ISSN 00223808. URL <http://www.jstor.org/stable/10.1086/341871>.
- João Ricardo Faria. The game academics play: Editors versus authors. *Bulletin of Economic Research*, 57(1):1–12, 2005. ISSN 1467-8586. doi: 10.1111/j.1467-8586.2005.00212.x. URL <http://dx.doi.org/10.1111/j.1467-8586.2005.00212.x>.
- Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, Oxford, 2007.
- Robert D. Gordon. Values of Mills’ ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941. ISSN 00034851. URL <http://www.jstor.org/stable/2235868>.
- Remco Heesen. Academic superstars: Competent or lucky? *Synthese*, forthcoming. doi: 10.1007/s11229-016-1146-5. URL <http://dx.doi.org/10.1007/s11229-016-1146-5>.
- Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, second edition, 1994.
- Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, Oxford, 1993. ISBN 0195046285.

David N. Laband. Publishing favoritism: A critique of department rankings based on quantitative publishing performance. *Southern Economic Journal*, 52(2):510–515, 1985. ISSN 00384038. URL <http://www.jstor.org/stable/1059636>.

David N. Laband and Michael J. Piette. Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy*, 102(1):194–203, 1994. ISSN 00223808. URL <http://www.jstor.org/stable/2138799>.

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>.

D. Lindsey. Using citation counts as a measure of quality in science: Measuring what’s measurable rather than what’s valid. *Scientometrics*, 15(3–4):189–203, 1989. ISSN 0138-9130. doi: 10.1007/BF02017198. URL <http://dx.doi.org/10.1007/BF02017198>.

Christopher D. Mackie. *Canonizing Economic Theory: How Theories and Ideas Are Selected in Economics*. M. E. Sharpe, New York, 1998. ISBN 9780765602848.

Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks. The independence thesis: When individual and social epistemology diverge. *Philosophy of Science*, 78(4):653–677, 2011. ISSN 00318248. URL <http://www.jstor.org/stable/10.1086/661777>.

Marshall H. Medoff. Editorial favoritism in economics? *Southern Economic Journal*, 70(2):425–434, 2003. ISSN 00384038. URL <http://www.jstor.org/stable/3648979>.

- 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. *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago, 1973. ISBN 0226520919.
- Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. Science faculty’s subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479, 2012. doi: 10.1073/pnas.1211286109. URL <http://www.pnas.org/content/109/41/16474.abstract>.
- Michael J. Piette and Kevin L. Ross. A study of the publication of scholarly output in economics journals. *Eastern Economic Journal*, 18(4):429–436, 1992. ISSN 00945056. URL <http://www.jstor.org/stable/40325474>.
- 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.
- Daniel L. Sherrell, Joseph F. Hair, Jr., and Mitch Griffin. Marketing academicians’ perceptions of ethical research and publishing behavior. *Journal of the Academy of Marketing Science*, 17(4):315–324, 1989. ISSN 0092-0703. doi: 10.1007/BF02726642. URL <http://dx.doi.org/10.1007/BF02726642>.
- Kenneth J. Smith and Robert F. Dombrowski. An examination of the relationship between author-editor connections and subsequent citations of auditing research articles. *Journal of Accounting Education*, 16(3–4):497–506, 1998. ISSN 0748-5751. doi: 10.1016/S0748-5751(98)

00019-0. URL <http://www.sciencedirect.com/science/article/pii/S0748575198000190>.

Rhea E. Steinpreis, Katie A. Anders, and Dawn Ritzke. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 41(7–8):509–528, 1999. ISSN 0360-0025. doi: 10.1023/A:1018839203698. URL <http://dx.doi.org/10.1023/A:1018839203698>.

Brandon D. Stewart and B. Keith Payne. Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, 34(10):1332–1345, 2008. doi: 10.1177/0146167208321269. URL <http://psp.sagepub.com/content/34/10/1332.abstract>.

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>.

Eric Luis Uhlmann and Geoffrey L. Cohen. “I think it, therefore it’s true”: Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104(2):207–223, 2007. ISSN 0749-5978. doi: 10.1016/j.obhdp.2007.07.001. URL <http://www.sciencedirect.com/science/article/pii/S0749597807000611>.

Virginia Valian. *Why So Slow? The Advancement of Women*. MIT Press, Cambridge, 1999. ISBN 9780262720311.

Christine Wennerås and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387(6631):341–343, May 1997. ISSN 0028-0836. doi: 10.1038/387341a0. URL <http://dx.doi.org/10.1038/387341a0>.

Alan J. Ziobrowski and Karen M. Gibler. Factors academic real estate authors consider when choosing where to submit a manuscript for pub-

lication. *Journal of Real Estate Practice and Education*, 3(1):43–54, 2000. ISSN 1521-4842. URL <http://ares.metapress.com/content/1762151051KM2227>.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6:185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://journals.cambridge.org/article_S1742360000001283.

Strategies of Explanatory Abstraction in Molecular Systems Biology[†]

Nicholaos Jones[‡]

Abstract

I consider three explanatory strategies from recent systems biology that are driven by mathematics as much as mechanistic detail. Analysis of differential equations drives the first strategy; topological analysis of network motifs drives the second; mathematical theorems from control engineering drive the third. I also distinguish three abstraction types: aggregations, which simplify by condensing information; generalizations, which simplify by generalizing information; and structurations, which simplify by contextualizing information. Using a common explanandum as reference point—namely, the robust perfect adaptation of chemotaxis in *Escherichia coli*—I argue that each strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details.

1 Introductory Remarks

The currently dominant paradigm for understanding explanation in biology puts mechanism at center stage (Nicholson 2012; Levy 2013). Leading accounts of mechanistic explanation, while differing in the particulars of their analysis of *mechanism*, agree that mechanistic explanations explain by alluding to mechanisms or models thereof (Machamer, Darden, Craver 2000; Bechtel and Abrahamsen 2005).

There is a small publishing industry devoted to discerning the scope of mechanistic explanation in scientific practice. Some claim to identify biological explanations that do not allude to mechanisms (Wouters 2007; Huneman 2010; Rice 2015). Fans of mechanistic explanation tend to resist making scope concessions, preferring instead to accommodate the putative explanations as mechanistic despite initial appearances, to broaden the scope of mechanistic explanation or the analysis of *mechanism*, or else to

[†] Draft. For symposium on *Integrating Explanatory Strategies Across the Life Sciences* at the 2016 meeting of Philosophy of Science Association, Atlanta, GA. I thank audiences at Mississippi State University, the Alabama Philosophical Society, and the Society for Philosophy of Science in Practice for comments on earlier drafts.

[‡] Department of Philosophy, University of Alabama in Huntsville, Huntsville AL 35899, nick.jones@uah.edu

deny that the putative explanations are explanations at all (Craver 2006; Bechtel and Abrahamsen 2010; Brigandt 2013; Levy and Bechtel 2013).

I set aside questions about what qualifies as an explanation as well as questions about whether only mechanisms—or models thereof—carry explanatory power. I focus, instead, on *explanatory strategies*, understood as patterns of reasoning directed toward providing explanations. I consider three explanatory strategies from recent systems biology that are driven by mathematics as much as, if not more than, mechanistic detail. Analysis of differential equations drives the first strategy; topological analysis of network motifs drives the second; mathematical theorems from control engineering drive the third.

Systems biologists use these strategies to supplement the explanatory power of traditional molecular mechanisms (see Brigandt et al *forthcoming*). My aim is to identify how the strategies differ from each other, rather than how they differ from standard mechanistic explanations or what might unify them in those differences (for which see Green and Jones 2016). Doing so helps with understanding relations among the strategies, their tactics for integrating mechanistic detail, and explanatory affordances of their mathematical elements.

The key to my analysis is a distinction among three abstraction types: aggregations, which simplify by condensing information; generalizations, which simplify by generalizing information; and structurations, which simplify by contextualizing information. Using a common explanandum as reference point—namely, the robust perfect adaptation of chemotaxis in *Escherichia coli* (Barkai and Leibler 1997; Ma et al 2009; Yi et al 2000)—I argue that each strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details. I begin with the typology of abstraction.

2 Abstraction Typology

I am interested in abstractions as representational rather than metaphysical. Abstractions, as I understand them, are ontologically innocent, so that characterizing features of representations as abstractions over some parts of reality carries no implication that features correspond to abstract objects (see also Cartwright 1989, 353-354; Levy and Bechtel 2013, 243). So, for example, representing the relation between a person, a hotel, and a date range as a reservation does not entail that some abstract object, a *reservation*, exists; nor does representing the motions of an object's constituents as the motion of the object's center of mass entail that some abstract object, a *center of mass*, exists.

Levy and Bechtel characterize a representation as abstract insofar as a more concrete representation is possible (2013, 242). Brigandt and colleagues suggest that biologists use abstractions to “elucidate system-level patterns of organization that may not be visible at the level of molecular details” (*forthcoming*). I concur. I understand abstractions as representing only some of the many elements—objects, relations, parameters—associated with their targets, thereby making apparent patterns obscured by more detailed representations. I add to these insights that biologists produce (at least) three types of abstraction.

Following Ordorica, I call the first *aggregation* (2015, 163-164). An aggregation represents some relationship among multiple elements of a representational target as a higher-level object, or multiple elements of the target as a single, composite object. (See Figure 1a.) Paradigm cases of aggregations include representations of person-hotel-date relations as *reservations*; of costs of services and costs of goods as *costs*; and of the motions of an object’s parts as the *motion of a center of mass* (from Ordorica 2015, 164). Aggregations abstract from plurality to individual, ignoring differences among many in order to make salient some integrated unity among the elements of a representational target. They thereby simplify representations by condensing information about representational targets.

Following Pincock, I call the second abstraction type *generalization* (2015, 864). A generalization represents some element of a representational target as a class of elements, where potential instances of the class might include elements not present in the target. (See Figure 1b.) For example, because the class of solution measures includes all soap-bubble-like surfaces, such as the cellular froth surrounding radiolarian protozoa, representing a soap-bubble surface as a “solution measure” is a generalization (Pincock 2015, 864). Generalizations abstract from an instance to a class thereof, ignoring differences between instances of the class in order to make salient some more general unity. They thereby simplify representations by generalizing from information about representational targets.

I call the third abstraction type *structuration*. A structuration represents some element of a representational target as a position in a structure, such that potential occupants of the position might include elements not present in the target. (See Figure 1c.) I follow Haslanger in understanding structures as “complex entities with parts whose behavior is constrained by their relation to other parts” (2016, 118). Paradigm cases of structururations include representating Barack Obama as President of the United States of America, or representing Alneias as son of Anchises and Aphrodite. Structurations abstract to a position in a structure, from an occupant of the position, ignoring intrinsic features of the occupant unrelated to its position in order to make salient the

occupant's role relative to occupants of other positions in the same structure. They thereby simplify by contextualizing information about representational targets.

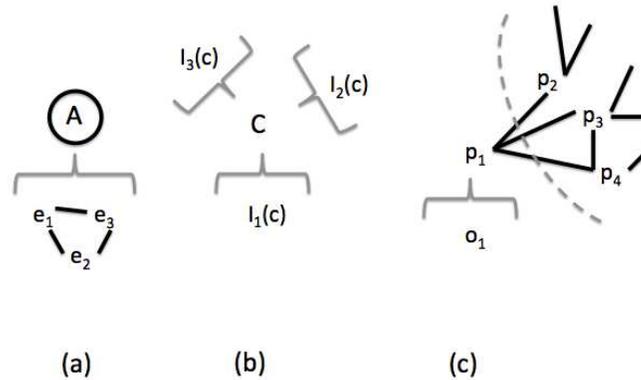


Figure 1: Visualizing Abstraction Types. (a) Aggregation A represents elements e_1 , e_2 , and e_3 (and relations therein) as a single object. (b) Generalization C represents $I_1(c)$ as a class, instances of which also include $I_2(c)$ and $I_3(c)$. (c) Structuration p_1 represents element o_1 as a position in larger structure that also includes p_2 , p_3 , and p_4 .

I understand aggregations as distinct from both generalizations and structurations, by virtue of being many-to-one, rather than one-to-one, simplifications. I also understand being a generalization as insufficient for being a structuration. For representations of positions carry information about functional relationships between their occupants and other positions in the same structure; but representations of classes do not. Finally, insofar as classes are sets, I understand being a structuration as insufficient for being a generalization. For, sometimes, representing target elements as classes carries some information about intrinsic features of those elements apart from their functional relations to elements occupying other positions in the same structure; but representing target elements as positions in structures never carries such information.

3 Robust Perfect Adaptation of *E. coli* Chemotaxis

My central claim is that different explanatory strategies from recent systems biology differ from each other, at least in part, by virtue of appealing to different abstraction types. I support this claim by considering a case in which multiple strategies target the same explanandum. Doing so minimizes confounds that confuse differences due to the nature of each explanatory strategy with differences due to the nature of each

explanatory target. I focus on a particular explanandum known as robust perfect adaptation of bacterial chemotaxis, following others who consider this a paradigmatic target for non-mechanistic explanation (Brallard 2010; Brigandt, Green, and O'Malley *forthcoming*; Matthiessen *forthcoming*).

3.1 Explanandum Context

Escherichia coli (*E. coli*) is popular model organism in biological research. It is very sensitive to small chemical changes over a very large range of background concentrations. It also has a simple and well-understood signal transduction network (Wadhams and Armitage 2004).

E. coli manages two kinds of motion (Berg 2003). It *runs* by rotating its flagellar motor counterclockwise. This aligns all of its flagella into a synchronized bundle, resulting in movement in a straight line for about 1 second. *E. coli* also *tumbles* by rotating its flagellar motor clockwise. This breaks flagellar alignment, and the asynchronized flagella produce stationary changes of direction lasting for about 0.1 second. *E. coli* are randomly reoriented after each tumble. Moreover, while these tumbles occur with regular frequency, *E. coli* with higher concentrations of CheR protein tumble more frequently (Spudich and Kochland 1975).

E. coli's motion in a uniform external environment resembles a random walk. *E. coli* has no ability to control or select its direction of motion, and its straight runs are subject to Brownian motion because of eddies. However, in the presence of a chemical attractant—amino acids such as serine or aspartic acid, or sugars such as maltose or glucose—*E. coli* *taxis* toward the attractant. This taxi behavior involves less frequent tumbles, leading to longer runs and so gradual motion toward the attractant. (There is an opposite behavior for repellants such as metal ions or leucine.)

The biomolecular mechanism for *E. coli* chemotaxis is well-understood. When an environmental attractant attaches to a receptor, the receptor lowers the activity of the CheW-CheA protein complex. Less activity from this complex reduces the rate of CheY phosphorylation, which results in less phosphorylated CheY diffusing to the flagella. Because CheY induces clockwise rotation of the flagellar motor, the outcome is less frequent tumbling.

3.2 Explanandum Question

Alon and colleagues have experimental verification that, in the presence of a chemical attractant mixed uniformly into the environment at a constant concentration, *E. coli* chemotaxis *perfectly adaptive* (Alon et al 2009). After a brief period of decreased tumbling frequency, the frequency of *E. coli* tumbles increases toward and returns to the

exact frequency prior to the introduction of the attractant. The effect of the attractant, accordingly, becomes entirely forgotten despite its continuing presence.

The biomolecular mechanism for the adaptiveness of chemotaxis for *E. coli* is also well-understood. Some time after a new attractant has been detected by receptors, the lower activity of the CheW-CheA complex induces less CheB activity. This reduces the rate for removing methyl groups from the CheW-CheA complex and, together with continual methylation of the CheR receptor, CheW-CheA methylation increases. More methylation means more CheW-CheA activity, which in turn induces more CheY phosphorylation. This eventually results in more phosphorylated CheY diffusing to the flagellar motor, which increases clockwise motor rotation and thereby raises tumbling frequency.

Alon and colleagues have further experimental verification that this perfectly adaptive chemotaxis of *E. coli* is *robust* across ranges of CheR concentrations 0.5 to 50 times higher than concentration levels in “wild type” *E. coli* (Alon et al 2009). (By contrast, *E. coli*'s adaptation time—the time to return to 50% of its pre-stimulus tumbling frequency—is not robust to different CheR concentrations, because more CheR entails longer adaptation times.) This is the explanandum of interest: why is the perfect adaptation of *E. coli* chemotaxis, in the presence of a well-distributed chemical attractant, robust to CheR protein concentrations?

There are (at least) three strategies for answering this question in recent systems biology literature. (For a fourth, see Kollman et al 2005.) I consider each in turn, first sketching the general strategy and then making explicit the abstractions at work.

4 Distinguishing Explanatory Strategies through Abstraction Types

4.1 Dynamical Modeling

I call the first strategy *dynamical modeling*. This strategy begins by constructing a chemotaxis network for *E. coli*. This network represents the mechanism for *E. coli* chemotaxis, including specific biochemical details about when and how relevant proteins affect each other. (See Figure 2.) For example, Barkai and Leibler (1997) construct a model according to which, among many other specifics, CheB demethylates only the active form of the CheW-CheA complex and CheR works only at saturation.

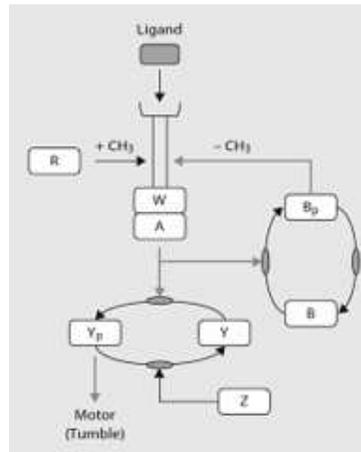


Figure 2. Mechanistic network for *E.coli* chemotaxis (Rao and Ordal 2009).

The dynamical modeling strategy proceeds by constructing a dynamical model—typically a set of differential equations—from the network (see Jones and Wolkenhauer 2012). One then demonstrates, via mathematical proof or simulation, that this model predicts perfect adaptation in the presence of a well-distributed chemical attractant for CheR concentration values varying over several orders of magnitude. (Raerinne 2013 calls this *sensitivity analysis*.) The demonstration supports the inference that *E.coli* chemotaxis exhibits robust perfect adaptation *because of its biochemical specifics*.

Bechtel and Abrahamsen (2010) call the product of this strategy a *dynamical mechanistic explanation*. I set aside the issue of whether the dynamical modeling strategy produces explanations. But I endorse Bechtel and Abrahamsen’s insight that the dynamical modeling strategy produces accounts that are mechanistic, by virtue of depending upon mechanistic details, as well as dynamical, by virtue of analyzing mathematical models built upon those details. For example, Barkai and Leibler’s (1997) mathematical analysis is relevant to *E.coli* chemotaxis only insofar as their network details are relevant; and analysis of the network apart from the model cannot produce an inference about the *robustness* of *E.coli*’s perfectly adaptive chemotaxis.

Let’s treat the dynamical model driving this explanatory strategy as an initial baseline for evaluating the number and severity of abstraction in various explanatory strategies. The model is abstract in various ways. But we shall treat it as a recipient of further abstractions, in the way a vehicle receives freight. Just as we can determine the weight of the freight indirectly by subtracting the gross weight of vehicle and freight from the “tare weight” (the weight of vehicle alone), we shall determine abstraction variety and

severity/extent for models driving other explanatory strategies by “subtracting” their total abstraction variety and severity from the “tare” abstraction.

4.2 Topological Analysis

I call the second explanatory strategy *topological analysis*. This strategy begins by identifying all possible minimal adaptation networks capable of predicting robust perfect adaptation for *E.coli* chemotaxis. These networks, like the networks for dynamical modeling, represent mechanisms for *E.coli* chemotaxis. Yet, unlike the networks for dynamical modeling, these networks are minimal: they contain the fewest possible nodes and links that suffice for robustly perfectly adaptive chemotaxis. The procedure for identifying all possible minimal networks of this sort is brute computational search. It turns out that there are exactly three, each of which has exactly three nodes and no more than three links (Ma et al 2009).

The topological analysis strategy proceeds by identifying a chemotaxis network known to predict robust perfect adaptation. This strategy thereby relies upon the dynamical modeling strategy, but only for mathematical results. The biochemical details of the chosen chemotaxis network turn out to be largely irrelevant, because the topological analysis strategy proceeds by demonstrating that a *reduced form* of the chosen network is topological equivalent to one of the minimal adaptation models. Reduced forms for mechanistic networks functional equivalents for node groups, group nodes or equivalents into modules, and ignore links within modules in favor of links between modules.

Consider, for example, one of the three minimal adaptation networks Ma and colleagues (2009) discover for *E.coli* chemotaxis. (See Figure 3.) The network has an input activating node A, A inhibiting being activated by B, A also activating C, and C activating some output. Ma and colleagues show that Barkai and Leibler’s (1997) model for *E.coli* chemotaxis reduces to this minimal network. Barkai and Leibler have an input and CheR activating, and CheB inhibiting, receptors; these receptors activating the CheW-CheA complex; the complex activating CheB and CheY; and CheY activating some output. Ma and colleagues reconceptualize Barkai and Leibler’s network into one where the input activates a *receptor complex*; this complex activates CheY, which activates the output; the complex also activates CheB, which inhibits a *methylation level* also activated by CheR.; and this methylation level activates the receptor complex. Then, in a second reconceptualization that produces one of their minimal adaptation networks, they group the receptor complex and CheB into module A, group CheR and the methylation level into module B, and rename CheY module C.

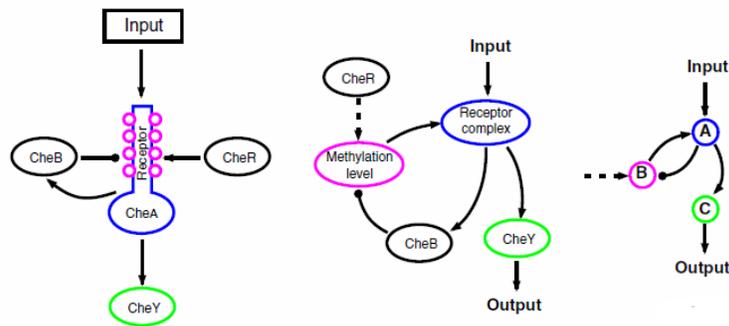


Figure 7. The Network of Perfect Adaptation in *E. coli* Chemotaxis Belongs to the NFBLB Class of Adaptive Circuits
 Left: the original network in *E. coli*. Middle: the redrawn network to highlight the role and the control of the key node "Methylation Level." Right: one of the minimal adaptation networks in our study.

Figure 3: Network topology for *E. coli* chemotaxis (Ma et al 2009).

The topological analysis strategy infers, from the topological equivalence between a minimal adaptation network and the reduced form of a network known to predict robust perfect adaptation for chemotaxis, that *E. coli* chemotaxis exhibits robust perfect adaptation *because of the topology of its chemotaxis network*. Huneman (2010) calls the product of this strategy a *topological explanation*. Regardless of whether analyses such as Ma and colleagues's are explanatory, they are topological by virtue of demonstrating some consequence about the topological properties of a network. This means that, even if the mechanistic details of *E. coli*'s chemotaxis network were different, and even if the biochemical specifics of the network chosen for reduction were different, the product of the topological analysis strategy would remain the same provided that the alternative networks preserve topological equivalence with the originals (see also Jones 2014).

The topological model driving this second explanatory strategy is more abstract than the dynamical model driving our initial ("tare") strategy. The topological model contains more aggregations. For example, it represents CheY and CheZ as "the motor rotation group;" it represents CheA and CheW as "the receptor complex;" and it represents the receptor complex and CheB as "the phosphorylation group." The topological model also contains more structurations. For example, it represents the phosphorylation group as "A" and the motor rotation group as "C." These representations abstract entirely from any intrinsic marks that might distinguish instances of "A" from instances of "C," relying instead upon extrinsic relations to distinguish the nodes from each other. So, for example, "A" but not "C" inhibits "B," "A" activates "C," and so on.

4.3 Organizational Design

I call the third explanatory strategy *organization design*. This strategy begins with a proof to the effect that systems exhibit robust perfect adaptation if and only if they

satisfy the characteristic equation for Integral Feedback Control (IFC). The proof is purely mathematical, well-known from control engineering theory in contexts involving mechanical systems that exhibit IFC such as thermostats. I am not aware of a complete and published version of this proof, but Yi and colleagues (2000) provide a sketch with relevant details. The organizational design strategy proceeds by inferring that *E.coli* chemotaxis exhibits robust perfect adaptation if and only if it satisfies the characteristic equation for IFC, and further inferring that *E.coli* chemotaxis exhibits robust perfect adaptation *because it satisfies the characteristic equation for IFC*. (For better explanatory details regarding this specific case, Braillard 2010; Green and Jones 2016.)

The organizational design strategy invokes neither mechanistic specifics about the chemotaxis network for *E.coli* nor topological details about the structure of that network. The strategy takes the explanandum phenomenon as given, using a mathematical equivalence result to identify a principle both necessary and sufficient for the phenomenon. The strategy thereby has affinities with explanatory strategies that appeal to organizing principles (Green and Wolkenhauer 2013) and design principles (Green 2015).

For simplicity, let's "reset" our abstraction "tare" to the topological model, because the model driving the organizational design strategy—call it the design model—is abstract in all the ways the topological model is abstract and more besides. The simplification thereby focuses attention on ways in which the design model differs from the topological model—and, by extension, from the initial dynamical model.

Compared to the topological model, the design model contains more aggregations. For example, the design model represents CheY phosphorylation and CheB activation as "k-box output." This aggregation is, at the same time, a generalization and a structuration. For example, "k-box output" is a class, with instances biological as well as mechanical. The standard example of a mechanical instance is heater activation in a thermostat. The k-box representation is also a structuration, akin to the "A", "B," and "C" representations from the topological model. For the k-box represents whatever has such-and-such input and output (a position in a structure). (See Figure 4.)

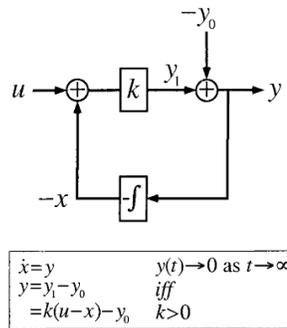


Fig. 2. A block diagram of integral feedback control. The variable u is the input for a process with gain k . The difference between the actual output y_1 and the steady-state output y_0 represents the normalized output or error, y . Integral control arises through the feedback loop in which the time integral of y , x , is fed back into the system. As a result, we have $\dot{x} = y$ and $y = 0$ at steady-state for all u . In the Barkai-Leibler model of the bacterial chemotaxis signaling system, the chemoattractant is the input, receptor activity is the output, and $-x$ approximates the methylation level of the receptors.

Figure 4 Organizational design for bacterial chemotaxis...and thermostats (Yi et al 2000).

The topological model is more abstract than the dynamical model, by virtue of containing various abstractions over protein identities. The design model, in turn, is more abstract than the topological and dynamical models, by virtue of also containing various abstractions over protein interactions. We can, therefore, arrange the various explanatory strategies along a continuum of abstraction type and severity. The dynamical modeling strategy, as our baseline, occupies the “low” end of our continuum. Next is topological analysis, which involves aggregations of and structurations from protein identities (or aggregations thereof). Then there is organizational design, which also involves aggregation of protein interactions as well as generalization and structuration of protein identities (or aggregations thereof).

5 Confirming the Analysis

I consider the foregoing to establish that each explanatory strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details. Whether this result generalizes beyond my chosen case study awaits future research. There is some reason to expect an affirmative result. For if dynamical, topological, and design explanatory strategies differ as I claim—specifically, along dimensions of number and severity of generalizations and structurations—then we should expect the *more abstract* strategies to have *wider scope*. For the more general models likely have more instances, and the more structural models likely have more position occupants.

We find confirmation of this prediction for the case of robust perfect adaptation of *Bacillus subtilis* (*B.subtilis*) chemotaxis. Details of the organization design strategy for explaining why *E.coli* chemotaxis exhibits robust perfect adaptation *also* apply for explaining why *B.subtilis* chemotaxis exhibits robust perfect adaptation. But details of the corresponding dynamical mechanistic strategy do not. The organization design strategy, as we know, involves more generalization and structuration than the dynamical mechanistic strategy. This confirms our prediction.

Allow me to be brief with the details. Rao and Ordal (2007) develop a dynamic mechanistic explanation for the perfect robustness of chemotaxis for *B.subtilis*. Their explanatory strategy follows the same pattern as Barkai and Leibler's in the case of *E.coli*. But details differ. For example, according to Barkai and Leibler's model, CheB in *E.coli* demethylates only active receptor complexes; according to Rao and Ordal, CheB in *B.subtilis* demethylates inactive ones too. Again, according to Barkai and Leibler's model, without CheY *E.coli* runs but does not tumble; according to Rao and Ordal, without CheY *B.subtilis* tumbles but does not run. One more: according to Barkai and Leibler's model, *E.coli* without CheB cannot run; according to Rao and Ordal, *B.subtilis* without CheB can run. See Figure 5.

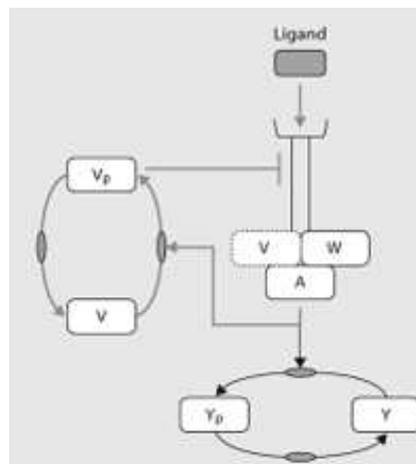


Figure 5: Chemotaxis network for *B.subtilis* (Rao and Ordal 2009).

So Barkai and Leibler's dynamical mechanistic explanation does not apply for the case of *B.subtilis*. But Yi and colleague's organizational design strategy does. For *B.subtilis*, like *E.coli*, exhibits robust perfect adaptation for chemotaxis if and only if it satisfies the characteristic equation for integral feedback control.

6 Toward Abstractive Mechanistic Explanation and its Affordances

Systems biological strategies for explaining the robust perfect adaptation of bacterial chemotaxis (in *E.coli*, *B.subtilis*, etc) apply mathematical techniques to network models. Dynamical, topological, and design strategies apply different techniques to explain the same phenomenon. Each explanatory strategy, moreover, applies its mathematical techniques to network models that embody different kinds and severities of these abstractions such as aggregations, generalizations, structurations. These abstraction types, accordingly, help to explain how these systems biological explanatory strategies differ from each other.

These abstraction types also provide a foundation for unifying various explanatory strategies from systems biology under the banner of mechanistic explanation. Let's consider well known kinds of mechanistic explanation as *standard*. Let's also follow Bechtel and Abrahamsen (2010) by considering *dynamical* mechanistic explanation as a mathematized species of standard mechanistic explanation.

Then let an *abstract network* be any network representation obtained by aggregating, generalizing, or structuring mechanistic details of the sort familiar in standard mechanistic explanation. Also let an *abstractive* mechanistic explanation be any explanation driven by applying mathematical techniques to an abstract network. See Figure 6.

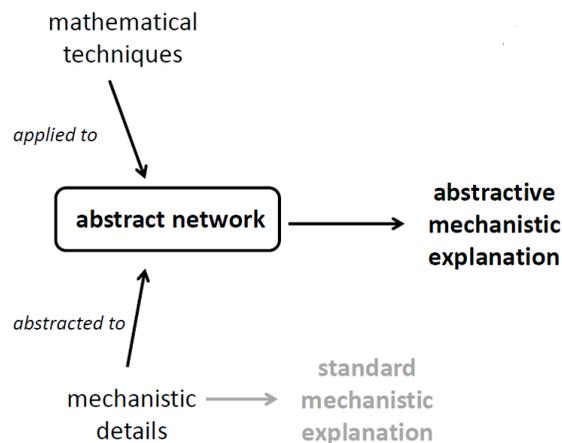


Figure 6. Relating standard and abstractive mechanistic explanation.

Then topological and organizational design explanatory strategies are mechanistic strategies—albeit abstractive ones. Topological explanations apply topological analysis

to aggregated and generalized mechanism networks. Organizational design explanations apply control systems engineering to aggregated, generalized, and structured mechanism networks.

Both kinds of explanation are mechanistic, by virtue of being grounded upon mechanistic details. But both also provide explanatory affordances unavailable through standard mechanistic explanations, by virtue of being abstract. For example, by virtue of using generalizations, topological explanations should have a greater scope than their standard mechanistic counterparts. By virtue of using generalizations and structurations, organizational design explanations should have still greater scope.

That these abstractive mechanistic strategies use novel mathematical techniques is a side effect of their using novel abstractions (in comparison with standard mechanistic explanations and their dynamical cousins). These techniques, of course, support more general conclusions, with wider scope, than the kind of differential equation analysis available for dynamical mechanistic explanations. But the techniques do not explain why the strategies have broader scope.

References

- U.Alon, M.G.Surette, N.Barkai, and S.Leibler, "Robustness in Bacterial Chemotaxis," *Nature* 397 (2009), 168-171.
- N.Barkai and S. Leibler. "Robustness in simple biochemical networks," *Nature* 387 (1997), 913-917.
- W.Bechtel and A.Abrahamsen, "Explanation: A Mechanistic Alternative," *Studies in History and Philosophy of Biological and Biomedical Science* 36 (2005), 421-441.
- W.Bechtel and A.Abrahamsen, "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science," *Studies in History and Philosophy of Science* 41 (2010), 321-333.
- H.C.Berg, *E.coli in Motion* (Springer, 2003).
- P.A. Braillard, "Systems Biology and the Mechanistic Framework," *History and Philosophy of the Life Sciences* 32.1 (2010), 43-62.
- I.Brigandt, "Systems Biology and the Integration of Mechanistic Explanation and Mathematical Explanation," *Studies in History and Philosophy of Biological and Biomedical Sciences* 44 (2013), 477-492.
- I.Brigandt, S.Green, and M.O'Malley, "Systems Biology and Mechanistic Explanation," in S. Glennan and P. Illari (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (forthcoming).
- N.Cartwright, "Capacities and Abstraction," in P. Kitcher and W. Salmon (eds.), *Scientific Explanation* (University of Minnesota Press, 1989), 349-356.
- C.F.Craver, "When Mechanistic Models Explain," *Synthese* 153 (2006), 355-376.

- S.Green, "Revisiting Generality in Biology: Systems Biology and the Quest for Design Principles," *Biology and Philosophy* 30.5 (2015), 629-652.
- S.Green and N.Jones, "Constraint-Based Reasoning for Search and Explanation: Strategies for Understanding Variation and Patterns in Biology," *dialectica* 70.3 (2006), 343-374.
- S.Green and O.Wolkenhauer, "Tracing Organizing Principles: Learning from the History of Systems Biology," *History and Philosophy of the Life Sciences* 35 (2013), 553-576.
- S. Haslinger, "What is a (Social) Structural Explanation?" *Philosophical Studies* 173 (2016), 113-130.
- P.Huneman, "Topological Explanation and Robustness in Biological Sciences," *Synthese* 177 (2010), 213-245.
- N.Jones, "Bowtie Structures, Pathway Diagrams, and Topological Explanation," *Erkenntnis* 79.5 (2014), 1135-1155.
- N.Jones and O.Wolkenhauer, "Diagrams as Locality Aids for Search and Explanation in Molecular Cell Biology," *Biology and Philosophy* 27 (2012), 1135-1155.
- M.Kollman, L.Løvdo, K.Bartholome, J.Timmer, and V.Sourjik, "Design Principles of a Bacterial Signalling Network," *Nature Letters* 438.24 (2005), 504-507.
- A.Levy, "Three New Kinds of Mechanism," *Biology and Philosophy* 28.1 (2013), 99-114.
- A.Levy and W.Bechtel, "Abstraction and the Organization of Mechanisms," *Philosophy of Science* 80 (2013), 241-261.
- W.Ma, A.Trusina, H.El-Samad, W.A.Lim, and C.Tang, "Defining Network Topologies that Can Achieve Biochemical Adaptation," *Cell* 138 (2009), 760-773.
- P.Machamer, Lindley Darden, and C.F.Craver, "Thinking about Mechanisms," *Philosophy of Science* 67 (2000), 1-25.
- D.Matthiessen, "Mechanistic Explanation in Systems Biology: Cellular Networks," *British Journal for Philosophy of Science* (forthcoming).
- D.J.Nicholson, "The Concept of Mechanism in Biology," *Studies in History and Philosophy of Biological and Biomedical Sciences* 43.1 (2012), 152-163.
- S.A.G.Ordorica, "The Explanatory Role of Abstraction Processes in Models: The Case of Aggregations," *Studies in History and Philosophy of Science* (2015), 161-167.
- C. Pincock, "Abstract Explanations in Science," *British Journal for the Philosophy of Science* 66 (2015), 857-878
- J.Raerinne, "Robustness and Sensitivity of Biological Models," *Philosophical Studies* 166 (2013), 285-303.
- C.V.Rao and G.W.Ordal, "The Molecular Basis of Excitation and Adaptation during Chemotactic Sensory Transduction in Bacteria," in M. Collin and R. Schuch (eds.), *Bacterial Sensing and Signaling* (Karger: Basel, Switzerland, 2009), 33-64.
- C.Rice, "Moving Beyond Causes: Optimality Models and Scientific Explanation," *Nous* 49 (2015), 589-615.

- J.L.Spudich and D.E.Kochland, "Non-Genetic Individuality: Chance in the Single Cell," *Nature* 262 (1976), 467-471.
- G.H.Wadhams and J.P.Armitage, "Making Sense of It All: Bacterial Chemotaxis," *Nature Reviews: Molecular Cell Biology* 5 (2004), 1024-1037.
- A.G.Wouters, "Design Explanation: Determining the Constraints on What Can Be Alive," *Erkenntnis* 67 (2007), 65-80.
- T.-M.Yi, Y.Huang, M.I.Simon, and J.Doyle, "Robust Perfect Adaptation in Bacterial Chemotaxis through Integral Feedback Control," *PNAS* 97 (2000), 4649-4653.

How the Diachronic Theoretical Virtues Make an Epistemic Difference

Mike Keas • Professor of the History and Philosophy of Science • The College at Southwestern

Abstract. Among the virtues of good theories are those appropriately labeled diachronic: durability, fruitfulness, and applicability—the last of which is insufficiently recognized. Diachronic theoretical virtues *cannot* be instantiated in the original construction of a theory; subsequent development is required. By contrast, one *can* assess the degree to which a theory exhibits the following nine non-diachronic theoretical virtues in a theory's original construction: evidential accuracy, causal adequacy, explanatory depth, internal consistency, internal coherence, universal coherence, beauty, simplicity, and unification. The distinction between diachronic and non-diachronic virtues is important for understanding the role and epistemic standing of each theoretical virtue.

Keywords. Theoretical virtues, durability, fruitfulness, prediction, and science-technology relations.

1. Introduction. Theoretical virtues are the traits of a theory that show it is probably true or worth accepting. Although the identification, characterization, classification, and epistemic standing of theory virtues are debated by philosophers and by participants in specific theoretical disputes, many scholars agree that these virtues help us to infer which rival theory is the best explanation (Lipton 2004). The most widely accepted theories across the disciplines usually exhibit many of the same theoretical virtues listed below. Each virtue class contains at least three virtues that sequentially follow a repeating pattern of progressive disclosure or expansion. In another forthcoming essay (Keas 2017) I argue for this new systematization of the theoretical virtues. In the present essay I focus on the diachronic class of virtues in contrast with the non-diachronic virtues. One can assess the degree to which a theory exhibits the non-diachronic virtues from the time a theory is initially framed. However, no theory, in its original construction, can instantiate

the diachronic virtues: durability, fruitfulness, or applicability. These virtues are instantiable only as a theory is later refined or applied.

Evidential virtues

1. Evidential accuracy: A theory (T) fits the empirical evidence well (regardless of causal claims).
2. Causal adequacy: T's causal factors plausibly produce the effects (evidence) in need of explanation.
3. Explanatory depth: T excels in causal history depth or in other depth measures such as the range of counterfactual questions that its law-like generalizations answer regarding the item being explained.

Coherential virtues

4. Internal consistency: T's components are related to each other logically.
5. Internal coherence: T's components are coordinated into an intuitively plausible whole; T lacks ad hoc hypotheses—theoretical components merely tacked on to solve isolated problems.
6. Universal coherence: T sits well with (or is not obviously contrary to) other warranted beliefs.

Aesthetic virtues

7. Beauty: T evokes aesthetic pleasure in properly functioning and sufficiently informed persons.
8. Simplicity: T explains the *same facts* as rivals, but with *less* theoretical content.
9. Unification: T explains *more kinds of facts* than rivals with the *same* amount of theoretical content.

Diachronic virtues

10. Durability: T has survived testing by successful prediction or plausible accommodation of new data.
11. Fruitfulness: T has generated additional discovery by means such as successful novel prediction, unification, and non ad hoc theoretical elaboration.
12. Applicability: T has guided strategic action or control, such as in science-based technology.

We will survey the first nine virtues only to the brief extent needed to recognize how one can assess the degree to which a theory exhibits these theoretical virtues in its original construction. This will, by contrast, enable us to appreciate the unique temporal character of the diachronic theoretical virtues.

2. Non-Diachronic Theoretical Virtues. We begin with the first three virtues. *Evidential accuracy*, which is how well a theory fits the relevant data, can be assessed from the theory's original construction. Often a theory will also, from its inception, specify *causally adequate* mechanisms to produce the phenomena in question. Such is not necessarily the case, as Alfred Wegener's theory of continental drift illustrates. His theory enjoyed considerable evidential accuracy despite its lack of a plausible cause to move the continents. *Explanatory depth* is also instantiated in a theory's initial formulation if, for example, the

theory answers a large range of counterfactual questions about a kind of phenomenon using the resources of its law-like generalizations.

The remaining six non-diachronic theoretical virtues likewise can be exhibited in the initial formation of a theory. A theory may be constructed in a logical manner so as to produce *internal consistency*. Beyond that, the theoretical components might be well coordinated into an intuitively plausible whole (avoiding ad hoc hypotheses), thus generating the theoretical virtue of *internal coherence*. If the theory sits well with (or is not obviously contrary to) other warranted beliefs, then it possesses the virtue of *universal coherence*. A new theory might even evoke aesthetic pleasure in the minds of experts, which constitutes theoretical *beauty*. The closely related virtues of simplicity and unification also might be instantiated in the initial formation of a theory: explaining the same facts as rival theories but with less theoretical content (*simplicity*), and explaining more kinds of facts than rivals with the same amount of theoretical content (*unification*).

Much more could be said about the first nine virtues outlined above (Keas 2017), but this is sufficient to recognize them as a group of theoretical virtues that can, in principle, be instantiated in a theory's original formation. This common trait remains characteristic of these virtues even (largely) under the disparate accounts found in the literature of how to characterize each virtue. Let us now explore the chief diachronic theoretical virtues in contrast to the non-diachronic virtues.

3. Diachronic Theoretical Virtues. Durability, fruitfulness, and applicability, which I recognize as the chief diachronic theoretical virtues, can only be instantiated as a theory is cultivated *after* its origin. This necessarily extended temporal dimension of the diachronic virtues is, arguably, of considerable epistemic importance. But even if one endorses the arguments that discount the epistemic significance of this temporal component (Mayo 2014), one still should acknowledge a group of virtues that (unlike the other the-

oretical virtues) can only be instantiated in a theory *after* its initial formulation. Time is of their essence in a manner that goes beyond the trivial truth that all human endeavor is temporal. McMullin (2014) has led the way in articulating the epistemic significance of two of the three main diachronic virtues: durability and fruitfulness (I recognize McMullin's third diachronic virtue of "consilience" as a mode of fruitfulness). Applicability, largely overlooked as a theory virtue, is another important member of this diachronic category, as I shall demonstrate.

3.1. Durability. Durability, a virtue term McMullin (2014) recommended, refers to the favorable epistemic condition of a theory that has survived testing by successful prediction or by plausible accommodation of new unanticipated data (or both). Popular or long-lived theories are not necessarily durable in the epistemic sense in view here. Equating durability with popularity or tradition is fallacious. While testability is a pragmatically admirable trait of a theory, it is not an intrinsic epistemic characteristic of a theory; many testable theories have failed too many tests to be acceptable. Steel (2010, 18) notes that the "more precise and informative a theory's empirical predictions are, the greater its testability." The more testable a theory is, the more durable it would prove itself to be if it passes the tests. A theory that scores low in testability has little potential to exhibit durability.

Despite the leading role of predictive success in many areas of science, it is less prominent in some reputable scientific theories that are, nevertheless, well endowed with other virtues. Successful prediction is very frequently part of explaining "how things work," but less routine in explaining "how things originated"—as in theories about the history of the cosmos, earth, and life (Cleland 2011, but Winther 2009 argues otherwise). Successful historical theories typically enjoy other forms of durability, most notably a track record of plausible accommodation of new data that, although not predicted, came to light after the theory's origin. The durability of a theory suffers if one or more of its predictions are disconfirmed

or when theorists respond to disconfirming evidence by modifying the theory with ad hoc hypotheses— theoretical components merely “tacked on” to solve isolated problems. Although initially a theory may exhibit a high degree of evidential accuracy (or any other of the first nine virtues in my systematization), it is impossible for a newborn theory to instantiate the virtue of durability—this *takes time* in a sense not required by the non-diachronic virtues. A similar necessary temporal dimension characterizes fruitfulness.

3.2. *Fruitfulness.* Fruitfulness, also known as fertility or fecundity, is another diachronic theoretical virtue. A theory is fruitful if, over time, it generates additional discovery by means such as successful novel prediction, unification, and non ad hoc theoretical elaboration. While durability is about conservation (a theory passing tests to survive), fruitfulness is about innovation (a theory stimulating further discovery). When a prediction formulated in the context of a theory’s construction is later verified, this successful predictive outcome increases the virtue of durability in that theory. By contrast, a *novel* prediction is one that was not conceived in conjunction with a theory’s construction, but that nevertheless follows reasonably from it. When such a novel prediction is confirmed by observation, a theory exhibits more fruitfulness.

The closely related diachronic character of durability and fruitfulness is well illustrated in the discovery of the first two planets beyond Saturn. Soon after Friedrich William Herschel unexpectedly discovered Uranus in 1781, astronomers noted that its observed motion strayed from what contemporary Newtonian mechanics predicted of such a planet. However, given the overall theoretically virtuous status of Newtonian physics up through that time (including its durability due to its success in testing), most astronomers expected a forthcoming way to make Uranus compliant with established theory. Even rejecting the anomalous data as “inaccurate” seemed reasonable early on. By the 1830s, however, the possibility of a perturbing planet beyond Uranus became a more reasonable and popular speculation, despite the ab-

sence of a precise novel prediction of where to find such a planet. By this time many astronomers were modestly confident in the accumulated data of Uranus' positions in the sky.

This brings us to the celebrated successful novel prediction of 1845-1846. Based principally on Newtonian physics and the well-known irregularities in Uranus' motion, two astronomers independently predicted where another unknown perturbing planet (later called Neptune) was likely located. Le Verrier's estimate of the planet's location was the most accurate (correct within one degree), as confirmed by a German astronomer on September 23, 1846. The (*fruitful*) novel prediction of Neptune was born within the context of a *durable* Newtonian orbital mechanics research tradition and the unexpected discovery of Uranus with its anomalous motions. The sensational success of this novel prediction (the discovery of Neptune) also rendered Uranus a Newtonian-compliant planet—thus further vindicating earlier provisional toleration of Uranus' anomalies, a toleration that had been justified by yet earlier Newtonian durability and fruitfulness.

Smith's (2010; 2014) landmark study of gravity theory from Newton to the present further illuminates the durability and fruitfulness of this research tradition, and it includes the case histories of Uranus and Neptune. Smith was surprised that the principal kind of question being tested was not "Do the calculated motions [e.g., of Uranus] agree with the observed motions?" Rather it was: "Can robust physical sources compatible with Newtonian theory be found for each clear, systematic discrepancy between the calculated and the observed motions?" Neptune (as novelly predicted) turned out to be such a robust physical source. However scientists failed over a half century to find a robust (detectable) physical source for the Newtonian-defying behavior of Mercury—a tiny anomaly in the precession of its perihelion. But this failure, which Einstein solved by way of theory replacement, does not completely diminish the enduring epistemic significance of two centuries of Newtonian durability and fruitfulness, as Hanson (1962) inaccurately suggested. Smith notes: "All the other discrepancies ended up revealing some detail of our plane-

tary system, the least subtle of which was Neptune, that theretofore had not been taken into account in the calculations” (2010, 552).

Such serial Newtonian problem solving became (almost always) ever more empirically constrained in a spiral of upward progress. For example, Uranus’ temporarily Newtonian-defying behavior “would have been masked if the significantly larger gravitational effects of Saturn on Uranus had not been included in the calculation first.” Smith explains further:

So, the discovery of Neptune provided evidence not only for Newton’s theory, but also for the specific aspects of Saturn that entered into calculating its effects on Uranus, for these were no less presupposed in the anomaly that emerged than Newton’s theory was. The point generalizes. Each time a discrepancy emerges and a robust physical source for it is found, that source is incorporated into the new calculations, and the process is repeated, typically with still smaller discrepancies emerging that were often theretofore masked in the calculations. So, what was being tested each time when a new discrepancy emerged and a physical source for it was being sought was not only Newtonian theory, but also all the previously identified details that make a difference and the differences they were said to make without which the further systematic discrepancy would not have emerged. (2010, 552-53)

On display is an interlocking of durability (passing tests to survive) and fruitfulness (stimulating further discovery) that is supportive of scientific realism. “This shows that increasingly strong evidence was accruing to Newtonian theory over the first two hundred years of orbital research based on it,” Smith concludes. This point (with some qualification) extends even to Einstein’s theoretical innovation that was partly justified by the unruly perihelion of Mercury. Einstein’s achievement was, to some degree, a continuation of this same progressive spiral, as Smith deftly explains:

As is well known, Einstein required Newtonian gravitation to hold in an asymptotic limit as he developed his new theory of gravity—specifically in a static, weak-field limit. That he did so was just as well because the 43 arc-seconds per century anomaly in the perihelion of Mercury that was initially the sole evidence for his theory presupposes Newtonian gravity.... As a matter of historical fact, all of the details singled out as making detectable differences during the two centuries of prior research carried over intact into post-Einstein orbital mechanics. *Save for some qualifications concerning levels of precision, the same details are still making the same differences as before....* So, Newtonian theory must still have some sort of claim to being knowledge. (2010, 556-57)

Smith's continuity-of-knowledge claim invites comment. While much of the metaphysics associated with Newtonian theory has been repudiated, we nevertheless see an impressive degree of fruitful scientific continuity from Newtonian to modern physics (at least in the particular ways that Smith documents). In sum, Newtonian orbital mechanics enjoyed increasingly impressive interlocking durability and fruitfulness over multiple centuries, and its approximate legitimacy (not counting discarded Newtonian metaphysics) remains similarly well-grounded today under the revisionary umbrella of modern physics.

Though some philosophers have argued to the contrary (Collins 1994; Harker 2008), many scientists and philosophers think that predictive success—especially novel predictive success—is a stronger indicator of likely approximate truth than a theory's accommodation of data (Douglas and Magnus 2013). According to my systematization (which illuminates but does not settle this thorny issue), data accommodation refers to a theory's initial instantiation of the evidential virtues (evidential accuracy, causal adequacy, and explanatory depth), and a theory's subsequent instantiation of certain diachronic virtues, namely non-predictive durability (plausibly making sense of new unanticipated data) and non-predictive fruitfulness (especially non ad hoc theoretical elaboration that makes sense of new unanticipated data).

3.2.1 *Unification as a Mode of Fruitfulness.* Fruitful theory elaboration, whether by means of successful novel prediction or non ad hoc theoretical elaboration that makes sense of unanticipated evidence, often also makes sense of *new kinds* of data, and thus is additionally recognized as increasing a theory's unification. Earlier we encountered unification as a non-diachronic (aesthetic) theoretical virtue. The diachronic increase of unification differs somewhat from its non-diachronic cousin. The historian and philosopher of science William Whewell (1794–1866) called diachronic unification “consilience.” When a theory explains a new domain of facts in a surprising way, then it is fruitful in a consilient manner. McMullin writes in this regard:

A good theory will often display remarkable powers of unification, making different classes of phenomena “leap together” over the course of time. Domains previously thought to be disparate now become one, the textbook example, of course, being Maxwell's unification of magnetism, electricity, and light. Examples abound in recent science, a particularly striking one being the development of the plate-tectonic model in geology. Assuming that this unifying power manifests itself over time, it testifies to the epistemic resources of the original theory and hence to that theory's having been more than mere accommodation. (2014, 505)

McMullin contrasts diachronic unification with its non-diachronic counterpart: “If the unification was achieved by the original theory, however, the virtue involved would no longer be diachronic.” Instead, it would count (in my systematization) as an aesthetic theoretical virtue that I simply call “unification,” and that Lipton calls “variety” (and yet others call “broad scope”). Lipton favors the assumption that such “heterogeneous evidence provides more support than the same amount of very similar evidence” (Lipton 2004, 168). Despite my own inclination to accept Lipton's point, I recognize this as a somewhat debatable assumption about the epistemic significance of an aesthetic property. However, when unification increases

over time, especially by means of surprising convergences, then unification is less likely the result of the idiosyncratic aesthetic predispositions and clever accommodating skills of a theorist during theory formation. Thus fruitful diachronic unification has greater confirmatory power than a theory's initial degree of aesthetic unification.

3.2.2 The Role of Prediction in the Diachronic Virtues. Drawing from Douglas' work on the relationship of prediction to inferring the best explanation, I argue that predictive success (in the first two diachronic virtues explored above) extends the epistemic work of many non-diachronic theoretical virtues such as causal adequacy, explanatory depth, beauty, simplicity, and unification. These latter theory traits, which she collectively labels as "explanatory,"

appeal to us, not just because we are aesthetically driven creatures but because such virtues help us to use the explanation to think and, in particular, to think our way through to new predictions, new tests, new rigors for our beautiful explanation. (2009, 460)

Douglas also notes:

Predictions are valuable because they force us (when followed through) to test our theories, because they have the potential to expand our knowledge into new realms and because they hold out the possibility (if successful) of gaining some measure of control over natural processes. (2009, 455)

Transposing Douglas' insights into my taxonomic terms, predictions are valuable because they figure into all three of the major diachronic virtues: durability (testing theories successfully), fruitfulness (expanding "our knowledge into new realms"), and applicability (which includes "gaining some measure of control over natural processes"). Moreover, the operation of prediction ("saying before" at least in a logical if not

temporal sense) in these three theoretical virtues further supports my classification of them as diachronic. Lets us now explore the last major diachronic virtue of applicability.

3.3. Applicability. Applicability refers to when a theory is used to guide successful action (e.g., prepare for a natural disaster) or to enhance technological control (e.g., genetic engineering). High degrees of the virtue of applicability obtain when a theory that is used to guide such action or control provides more effective outcomes than what is possible in the absence of the theory. Successful scientific theories constitute *knowledge* of the world (knowing *that*), not *control* over the world (which is mainly knowing *how*) for practical (non-theoretical) purposes. In this regard Strevens (2008, 3) notes: “If science provides anything of intrinsic value, it is explanation. Prediction and control are useful ... but when science is pursued as an end rather than as a means, it is for the sake of understanding.” But even after the intrinsic good of a theoretically virtuous explanation is in hand, one of several possible additional confirmatory diachronic (predictive or controlling) virtues might be acquired by a theory, including applicability. In such cases a good theory just gets better—even more confidence in its probable truth is justified.

Although scientific experiments use technological control, they do so to test scientific theories—so the main function is still to understand nature, not to control it. However, especially in the case of theories supported by experimentally verified prediction, such foreknowledge and laboratory control might be exploited to achieve practical aims such as device fabrication or medical intervention. But in any case, one cannot *apply* scientific knowledge until *after* one first *obtains* it. This necessary time lapse makes applicability diachronic.

To obtain scientific knowledge we search for a theory that (initially) exhibits many of the non-diachronic theoretical virtues. Subsequent work aimed at theory testing and elaboration might produce the additionally confirming presence of the diachronic virtues of durability and fruitfulness. At some point in

this dance of virtue-driven theory assessment and refinement, sufficient confidence in a particular theory might spur attempts to apply it as the basis for a new or improved technology. If the derived science-based technology actually works, then the “applied theory” has acquired the additional theoretical virtue of applicability. Because this requires additional time after initial theory formation, the diachronic classification of applicability is appropriate.

Although the application of scientific theories constitutes one aspect of technology, most of technology involves the empirical discovery of “know how” knowledge without crucially presupposing or immediately applying any particular scientific theory. Indeed, the relation between science and technology is not a simple one-way linear affair (Radder 2009; Douglas 2014). But this “emancipation” of technology from subordination to science, accomplished by historians and philosophers of technology between 1960 and 1990 (Houkes 2009, 310), should not obscure the epistemic significance of instances of technological innovation made possible, in part, by *applied* scientific theory.

This point is in harmony with the so-called demise of the “pure vs. applied science” dichotomy. Understanding and controlling nature are closely related, as our study of the diachronic theoretical virtues, including applicability, indicates. Douglas (2014, 62) surfaces some of the subtlety of this argument when, on the one hand, she proclaims: “With the pure vs. applied distinction removed, scientific progress can be defined in terms of the increased capacity to predict, control, manipulate, and intervene in various contexts.” But then, on the other hand, in a footnote she recoils partially: “To be clear, while I think this is a useful rubric for scientific progress, it is not a remotely sufficient account for how one should assess scientific theories.” Other (non-diachronic) theoretical virtues that are complementary to, but less weighty epistemically than, prediction and control also play important roles in theory assessment, she suggests. Consideration of the nine major non-diachronic theoretical virtues systematized in Sections 1 and 2 drives this point home.

How exactly is applicability a diachronic theory trait that is *epistemic* (helping to indicate likely truth) in view of the obvious *pragmatic* orientation of technological application? Agazzi observes that some technological projects “are designed or projected in advance, as the concrete application of knowledge provided by a given science or set of sciences” (Agazzi 2014, 308). If a project of this kind actually works as predicted, then this reinforces our confidence in the theory base that helped guide such action in the world. Agazzi further notes:

The predictions ‘contained’ in the project actually are the predictions made by the scientific theories which have permitted the proposal of the complex *noema* that constitutes the project, and contains not only prescriptions as to the way of realising the structure of the machine but also as to its functioning. This functioning is something that happens; it is a state of affairs that constitutes a confirmation of the theories used in projecting the machine. (309)

Although Agazzi’s scientific realism overstates the epistemic reach of applicability, it is helpful nonetheless as a corrective to other philosophical errors:

A mature science is a science that has given rise to a significant technology. This means, for example, that we can provisionally admit certain theories that are ‘empirically adequate,’ without admitting their truth as van Fraassen says, until we have significant predictions confirming them. This fact (especially in conjunction with other ‘virtues’ discussed in the literature) already justifies attributing truth and ontological reference to them, but the existence of technological applications is the last decisive step that assures that they have been able to adequately treat those aspects of reality they intended to treat. These last words are very important. They underline the fact that technological success does not eliminate the partial or limited scope of scientific theories. The fact that we can use classical mechanics in creating many machines or for sending rockets into space certainly means that this mechanics is true of its

objects and therefore ‘tells a true story’ about certain aspects of reality. This can also be expressed by saying that this theory is partially true of reality, but only if we mean that it does not speak about the totality of the attributes of reality, and that, consequently, it can speak properly only of such referents that possess these attributes. In other words, it is not correct to say that this mechanics is true regarding the whole of reality because other aspects of reality exist that must be accounted for by means of other theories which, in turn, can be used as a basis for different technologies. (310-11)

To nuance Agazzi’s insightful but somewhat inflated epistemic role for applicability, we can observe that this theoretical virtue is not commonly operative in certain scientific domains. For example, scientific theories of “how things originated” (history of nature) lead to fewer technological applications than scientific theories of “how things work.” Part of the reason for the infrequent applicability of origins theories is the smaller role that experimentally controlled prediction plays in such theorization. For example, much of the data that allows us to reconstruct the *history* of earth’s surface is collected by means of passive field observations, rather than by laboratory experiments that make precise predictions and technological control more feasible.

4. Conclusion. The diachronic theoretical virtues possess a temporal dimension that is absent from the other theoretical virtues. They can only be instantiated *after* a theory’s initial formulation—when it has had opportunity to be tested, elaborated, and applied. Durability, fruitfulness, and applicability build upon the initial theory assessment process governed by the non-diachronic virtues (the evidential, coherential, and aesthetic theoretical virtues). The cumulative result, when successful, is a mature theory with an even greater probability of being true than an infant theory that has not yet had the opportunity to show whether it will possess the diachronic theoretical virtues (anti-realists are invited to interject their own alternative

to this realist understanding of the theoretical virtues). So, the distinction between diachronic and non-diachronic virtues is important for an adequate account of theory evaluation.

The three major diachronic theoretical virtues are also better understood when they are recognized as related to each other in the following progressive sequence. Durability is instantiated as a theory passes more rigorous tests in a series of encounters with the world, especially by successful prediction and plausible accommodation of new evidence. Fruitfulness discloses a theory's resourcefulness yet further through innovation—stimulating additional discovery by successful novel prediction, unification, non ad hoc theoretical elaboration, and other means. At last, applicability expands the epistemic accountability of a theory into the final frontier: the vast domain of practical action. This virtue is instantiated when a theory helps us to interact with the world successfully, most notably by technological control. Together, these diachronic theoretical virtues provide an ongoing and epistemically intensified means of theory development that complements the non-diachronic virtue assessment process that begins in a theory's original construction.

Applicability, as a theoretical virtue, has not received the attention it deserves. Surprisingly, it is absent from every theoretical virtue list I have encountered. My work sketches a way to understand applicability in relation to the other diachronic virtues, and the larger group of non-diachronic virtues. This endeavor promises to illuminate, among other things, discussion of realism vs. anti-realism, science-technology relations, and inference to the best explanation.

References

Agazzi, Evandro. 2014. *Scientific Objectivity and Its Contexts*. Cham: Springer.

Cleland, Carol E. 2011. "Prediction and Explanation in Historical Natural Science." *British Journal for the Philosophy of Science* 62 (3):551-582.

- Collins, Robin. 1994. "Against the Epistemic Value of Prediction over Accommodation." *Noûs* 28 (2):210-224.
- Douglas, Heather E. 2009. "Reintroducing Prediction to Explanation." *Philosophy of Science* 76 (4):444-463.
- 2013. "The Value of Cognitive Values." *Philosophy of Science* 80 (5):796-806.
- 2014. "Pure Science and the Problem of Progress." *Studies in History and Philosophy of Science Part A* 46:55-63.
- Douglas, Heather, and P. D. Magnus. 2013. "State of the Field: Why Novel Prediction Matters." *Studies in History and Philosophy of Science Part A* 44 (4):580-589.
- Hanson, Norwood Russell. 1962. "Leverrier: The Zenith and Nadir of Newtonian Mechanics." *Isis* 53 (3):359-378.
- Harker, David. 2008. "On the Predilections for Predictions." *The British Journal for the Philosophy of Science* 59 (3):429-453.
- Houkes, Wybo. 2009. "The Nature of Technological Knowledge." In *Philosophy of Technology and Engineering Sciences*, 309-350. Amsterdam: North-Holland.
- Lipton, Peter. 2004. *Inference to the Best Explanation*. 2nd ed. London: Routledge.
- Mayo, D. 2014. "Some Surprising Facts About (the Problem of) Surprising Facts." *Studies in History and Philosophy of Science Part A* 45:79-86.
- McMullin, Ernan. 2014. "The Virtues of a Good Theory." In *The Routledge Companion to Philosophy of Science*, ed. Martin Curd and Stathis Psillos, 561-571. New York: Routledge.
- Radder, Hans. 2009. "Science, Technology and the Science-Technology Relationship." In *Philosophy of Technology and Engineering Sciences*, ed. A. Meijers, 65-91. Amsterdam: North Holland.

- Smith, George E. 2010. "Revisiting Accepted Science: The Indispensability of the History of Science." *Monist* 93 (4):545-579.
- 2014. "Closing the Loop: Testing Newtonian Gravity, Then and Now." In *Newton and Empiricism*, ed. Zvi Biener and Eric Schliesser, 262-351. New York: Oxford University Press.
- Steel, Daniel. 2010. "Epistemic Values and the Argument from Inductive Risk." *Philosophy of Science* 77 (1):14-34.
- Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.
- Winther, R. G. (2009). Prediction in selectionist evolutionary theory. *Philosophy of Science*, 76(5), 889-901.

Reconciling axiomatic quantum field theory with cutoff-dependent particle physics

Adam Koberinski¹

¹Department of Philosophy, Western University

Abstract

The debate between Fraser and Wallace (2011) over the foundations of quantum field theory (QFT) has spawned increased focus on both the axiomatic and conventional formalisms. The debate has set the tone for future foundational analysis, and has forced philosophers to “pick a side”. The two are seen as competing research programs, and the major divide between the two manifests in how each handles renormalization. In this paper I argue that the terms set by the Fraser-Wallace debate are misleading. AQFT and CQFT should be viewed as complementary formalisms that start from the same physical basis. Further, the focus on cutoffs as demarcating the two approaches is also highly misleading. Though their methods differ, both axiomatic and conventional QFT seek to use the same physical principles to explain the same domain of phenomena.

1 Introduction

Foundational investigation into quantum field theory (QFT) has emerged as a flourishing enterprise in philosophy of science, thanks largely to work done in axiomatic QFT (AQFT), particularly the C^* -algebraic approach encoded by the Haag-Kastler axioms (Haag and Kastler 1964). Despite the methodological disconnect with ‘conventional’ approaches to QFT (CQFT), AQFT has been defended by Fraser (2009) as supplying a firmer foundation from which to conduct philosophical analyses. Though this is one of few explicit defenses of AQFT, the widespread use of algebraic methods in philosophical literature on QFT would lead one to believe that Fraser is merely making explicit the assumptions in her field. Recently, Wallace (2006; 2011) has questioned the focus on AQFT, arguing that CQFT is the better candidate for analysis. Since CQFT is the theory that has been empirically successful—the Standard Model of particle physics is built from CQFTs—and AQFT has yet to reproduce these results, Wallace argues that we should focus analysis on CQFT rather than AQFT. Fraser’s (2011) reply has set up what is now known as the Fraser-Wallace debate over the foundations of QFT. The debate has set the tone for future foundational analysis, and seems to force philosophers to “pick a side”—you either work in AQFT or CQFT. The two are seen as competing research programs, and the major divide between the two manifests in how each handles renormalization. AQFT requires strict Poincaré covariance at arbitrarily small length scales, while the renormalization group (RG) methods in CQFT allow for a small-scale cutoff, below which QFTs needn’t be well-defined.

In this paper I argue that the terms set by the Fraser-Wallace debate are misleading. One needn’t view AQFT and CQFT as rival research programs; in fact, this view is

detrimental to understanding the history and methodology of QFT. AQFT and CQFT should be viewed as complementary formalisms that start from the same physical basis. Further, the focus on cutoffs as demarcating the two approaches is also highly misleading: AQFT can accommodate cutoffs and RG methods, and CQFT does not explicitly require cutoffs. The focus on cutoffs as essential to CQFT could mistakenly be taken to mean that CQFT depends on cutoffs actually *being physical*, in the same way that cutoffs are physical in condensed matter physics (CMP). I will argue that this is not the case: cutoffs needn't be physical in any sense. Even if cutoffs are *physically significant*, that does not entail that the cutoffs are themselves physical. Specifically, RG methods provide no principled grounds for thinking that cutoffs are "real" in the sense of signifying a breakdown of field theories generally. Since Wallace (2011) set the terms of the debate, the bulk of the arguments in this paper will be in reference to that paper. I do not claim that Wallace holds all (or even most) of the views against which I argue; rather, I use his paper to clarify potential misconceptions that could arise from the debate. Renormalization is not central to the physical content of QFT, and the different ways of handling renormalization do not mark AQFT and CQFT as different research programs. We should instead view the formalisms as complementary: though their methods differ, both seek to use the same physical principles to explain the same domain of phenomena.

2 Renormalization and the relationship between AQFT and CQFT

Wallace (2011) emphasizes the ineliminable dependence on cutoffs in CQFT, along with the success of RG methods for providing a physical motivation for cutoffs, as the wedge which drives AQFT and CQFT apart. For Wallace, AQFT cannot deal with physical cutoffs. Since RG methods have physically legitimized cutoffs, AQFT and CQFT have differing physical content and must therefore be considered a different research program (2011, Sec. 2). I disagree with this characterization on two fronts. First, AQFT has the resources to incorporate RG methods when needed. Though typical axioms make no mention of scaling behaviour, even the most rigid of axiomatic approaches—algebraic QFT as codified in the Haag-Kastler axioms—can incorporate something like RG flows.¹ Second, the calculational dependence on cutoffs in CQFT may not signal the physical existence of cutoffs.

So, are cutoffs really that problematic for AQFT? Many axiomatic approaches to QFT make no recourse to cutoffs, either explicitly or implicitly. An explicit forbidding of cutoffs would mean that one of the axioms/postulates of the theory claimed that the theory is empirically adequate at all spacetime length scales. Even if any axiomatization contained such an axiom (none do), it would be hard to imagine what sort of work it would do in derivations. Presumably, such a system could be modified to remove the guilty axiom, without spoiling any physically useful theorems. One should therefore not be concerned with an explicit ban on cutoffs in AQFT.

The more interesting case is when cutoffs are implicitly rejected by a particular theory.

¹See Buchholz and Verch (1995) for an example of scaling algebras playing the role of RG flows.

There are two common assumptions in AQFT that are problematic for handling cutoffs: strongly continuous implementations of Lorentz invariance, and the association of algebras with arbitrarily small open bounded regions of spacetime. Though the latter is not common to all axiomatic QFTs (the Wightman axioms deal directly with quantum fields, rather than algebras), the dominant axiomatization in terms of C^* algebras—the Haag-Kastler axioms—define QFTs in terms of algebras of observables corresponding to open, bounded regions of spacetime.² It is implicit that for any open bounded spacetime region, *no matter how small*, one can define an algebra of observables satisfying the other axioms defining QFT. If cutoffs are physical, one might conclude that there should be a principled limit to the size of regions on which we can define algebras corresponding to observables in QFT. If the cutoff scale is physically relevant, and only CQFT predicts its existence, we might be tempted to conclude that the two are different, competing theories. However, there are several possibilities for reconciling AQFT and cutoffs, which I will outline below. These remedies are largely independent of one another, and organized in terms of increasing foundational disagreement with Wallace’s view of cutoffs. The “quick fixes” proposed first lead to further conceptual worries, and I therefore endorse the option in Sec. 2.3, which is the biggest departure from taking cutoffs as physical in CQFT. Nevertheless, all the options sketched below are more-or-less viable. Section 2.4 outlines reasons for thinking that *both* AQFT and CQFT suffer the same conceptual challenges if cutoffs *really are physical*.

²Since algebraic QFT is *prima facie* the most problematic, I will deal primarily with algebraic QFT in this paper. The reader can take AQFT to stand for axiomatic QFT or algebraic QFT for the remainder of this paper. The reader should also note that constructive QFT is another important strand of rigorous QFT. Though it is conceptually distinct from AQFT, the two projects often overlap.

2.1 Possibilities for cutoffs in AQFT

Just because we need to associate an algebra with any arbitrary open bounded region of spacetime, we are not therefore compelled to make this algebra interesting. One way that cutoffs could be introduced into AQFT is to specify that regions smaller than some 4-volume Λ are to be uniformly assigned trivial algebras, i.e., algebras containing only multiples of the identity. Such assignments would be consistent with the demand that all open bounded regions of spacetime be assigned an algebra, but it would make the cutoff physically relevant, since no information about local parameters would be contained in regions smaller than Λ .

Though this solution is available, it is admittedly somewhat ad hoc. Even worse, it violates one of the crucial Haag-Kastler axioms: that of weak additivity. The axiom of weak additivity states that, for *every* closed, bounded region \mathcal{O} of Minkowski spacetime \mathcal{M} , the C^* norm closure of the algebras $\mathfrak{A}(\mathcal{O} + \alpha)$ for $\alpha \in \mathbb{R}^4$ is just the quasilocal algebra for the whole spacetime, $\mathfrak{A}(\mathcal{M})$.³ There are two reasons why this is a problem for introducing cutoffs in the way described above. First, we run into the problem that the quasilocal algebra corresponding to the whole of \mathcal{M} can be constructed from *any* algebra corresponding to *any* closed, bounded region \mathcal{O} . The norm closure of extensions of a trivial algebra will not produce any interesting algebra as a result, so regions smaller than the cutoff Λ will violate weak additivity. Second, extensions of an arbitrary region \mathcal{O} by some $\alpha < \Lambda$ should not be physical if Minkowski spacetime breaks down at scales below Λ . In the spirit of the first ad hoc axiom modification, weak additivity could be modified to exclude regions $\mathcal{O}_{small} < \Lambda$, and arbitrary extensions $\alpha_{small} < \Lambda$. However,

³See Ruetsche (2011), especially chapters 4 and 5 for an introduction to algebraic QFT. For a more comprehensive review of algebraic QFT, see Halvorson and Müger (2007).

there seems to be no principled reason for choosing a specific value of Λ , and one may question the naturalness of such axioms. This makes the solution of simple axiom modification less tempting, and forces us to admit that AQFT—at least in its current guise—is in conflict with approaches to QFT that take cutoffs as physically meaningful, since the basic axioms are currently in direct conflict with the introduction of cutoffs. If we admit that there is currently no room in the formalism of AQFT for cutoffs, are we doomed to take AQFT as (incorrectly) positing its own validity at all energy scales?

2.2 No cutoffs? No problem

If QFT methods are only applicable up to some cutoff energy, and we expect QFT to incorporate this fact, we are saying that a good theory should signal its own demise. The formal necessity of cutoffs in the formalism of CQFT has led to the idea that our best theories will continue to be an increasing hierarchy of effective field theories. Each field theory requires cutoffs to be implemented at a certain energy scale, and this signals the field theory's domain of applicability. If supplanted by a successor field theory, one expects that the new theory's low energy regime reduces to the old theory, and further that the new theory will itself have a higher energy cutoff. Following this approach, the conventional formalism of field theories would allow us to climb higher and higher up the ladder of energy scales, but we would never reach the top. We would require a theory of a fundamentally different formal type in order to end the ladder of cutoffs. This is presumably the view that Wallace holds, as he claims that if we replace one field theory with another applicable at higher energies, “that field theory in turn will need some kind of short-distance cutoff” (2011, p. 118).

As great as it may be to have a framework in which theories limit their own domain of

applicability, this is certainly not a necessary condition that any good formalism need satisfy. Even if AQFT does not contain cutoffs explicitly, this does not make it at odds with CQFT. Many theories that have been useful in the past do not signal their ultimate demise; on the contrary, most are mathematically well-defined well beyond their domain of applicability. For example, classical theories of fluid dynamics treat fluids as classical continua, and these continua are uniform to arbitrary precision. Classical continuum fluid dynamics is a useful theory, and compatible with classical point mechanics, even though classical point mechanics leads one to believe that the continuum is only an approximation—at some point fluid dynamics must break down. There is nothing within the formalism of fluid mechanics that signals its eventual breakdown; rather, the physical systems we model using classical fluid dynamics, as well as the complementary formalism of classical point particles, give us a physical motivation for the eventual breakdown of the formalism. Deeper theories, such as quantum mechanics, also provide grounds for believing in the limited applicability of both of the complementary classical formalisms. Similarly, we can view AQFT as a complementary picture to the formalism of CQFT. Both formalisms rely on the same general physical principles, though they are implemented in different ways. Though the AQFT formalism does not demarcate its domain of applicability in the form of explicit cutoffs, the necessity of some form of cutoff in CQFT provides reason to believe that the AQFT formalism is only approximately mapping the actual physics. Further, whatever extratheoretical grounds we have for taking cutoffs to be physical—typically in the guise of speculative physics beyond the Standard Model—can inform the scale at which we lose faith in the predictions of *both* the AQFT and CQFT formalisms. When one does not view AQFT and CQFT as rival research programs, the two can work together to provide a deeper

physical understanding of high energy physics, and the role of cutoffs is made clearer.

2.3 Physical significance versus being physical

Are cutoffs really that central? The arguments in the previous section assume that the cutoffs required to generate predictions in CQFT are physical, in the sense that they signal a breakdown of QFT. The fact that perturbative calculations within a particular model diverge when the integrals are unbounded does not entail that field theoretic methodology loses physical significance near these bounds. Undoubtedly we have extratheoretical reasons for supposing that the QFTs making up the Standard Model are not accurate to arbitrary energies—at some point gravity will surely play an important role, to say nothing for possible unknown physics at higher energy scales—but this needn't signify a breakdown of QFTs *in general* beyond a cutoff. Nor is this notion built in to the conceptual apparatus of RG methods, as Wallace claims.⁴ It remains entirely possible that a QFT built with more terms in its Lagrangian could describe all relevant physics and be well-defined at all energy scales. In fact, the renormalization group procedure presupposes a theory given in terms of a Lagrangian or Hamiltonian with an arbitrary number of terms. These terms are shown to go to zero in the low energy limit (Wilson and Kogut 1974). We know—using the RG methods to determine the flow of coupling constants—that for non-Abelian gauge theories, interactions become weaker at higher energy scales. Total asymptotic freedom would be one way to eliminate cutoffs at

⁴“Wilson's explanation of the renormalisation procedure relies upon *the failure of the QFT to which it is applied* at very short distances. It is then intriguing to ask how to put on a firm conceptual footing a theory which relies for its mathematical consistency on its own eventual failure”. (Wallace 2006, 34, emphasis added) Again, this passage can be read in a way that agrees with the arguments of this section. I am attempting to argue against a naive reading, which takes the failure of *one* QFT (i.e., a single form of interaction, encoded in a particular Lagrangian) to signal the failure of QFT methods in general.

high energies. A successor QFT, such as a grand unified theory or supersymmetry, could therefore unite the strong and electroweak coupling constants, while remaining well-defined to arbitrarily high energies.⁵ All that RG methods rely on conceptually is the ability to average out behaviour at high energy scales, and this is compatible with many options for high-energy behaviour. First, our theories could be low-energy approximations that break down at higher energy scales. This could be due to a fundamental granularity or discreteness in the more fundamental theory, or due to the absence of terms in the Lagrangian modelling high energy dynamics. Second, we could have a well-defined high energy dynamics that is unimportant at the energy scales with which we are concerned. In any case, RG methods provide no principled grounds for thinking that cutoffs are “real” in the sense of signifying a breakdown of field theories generally. Unlike the breakdown of classical fluid mechanics—for which we have a more fundamental successor theory (quantum mechanics) providing grounds to reject the continuum as merely an approximation—there is as of yet no (empirically successful) fundamental successor theory for which QFT can be considered a continuum approximation.

One of the major reasons for thinking that cutoffs in QFT mark a regime beyond which the methods of QFT can no longer be applied is the success of RG methods originating from CMP (Wallace 2011, Sec. 1). RG methods were initially developed to investigate long range correlations in materials approaching a phase transition. Long range interactions are those most relevant to global transitions of a material, and so RG

⁵Whether a theory can be made well defined for arbitrarily high energies is a distinct issue from the accuracy of that theory’s predictions at high energies. It may turn out that Standard Model QFTs can be extended in a consistent way, but that the high energy predictions turn out to be false. This is the case that is argued in Section 2.2 regarding AQFT.

methods average out the unimportant short range behaviour near a critical point. The apparatus of non-relativistic QFT (i.e., functional integrals using Galilean invariant Lagrangians) is used in CMP as an *approximation* to the discrete atomic (or ionic) physical makeup of bulk systems. Given the the CMP field theories are explicitly constructed as approximations to a known underlying lattice model, we know that the field theoretic methods must break down within CMP. RG flow equations are derived by separating field variables φ into low- and high-momentum components $\varphi = \varphi_{low} + \varphi_{high}$ (where the cutoff from low to high is chosen arbitrarily) and averaging over the high momentum modes. The resulting Lagrangian $\mathcal{L}'(\varphi_{low})$ is then manipulated to fall into the same form as the original Lagrangian $\mathcal{L}(\varphi)$. This process is repeated and generates discrete recursive relations between the rescaled coupling parameters in the $(n + 1)$ th Lagrangian in terms of the n th one. In the limit where the rescalings are continuous, these become differential equations determining the flow of coupling constants under RG. As the flows are taken to zero frequency—equivalent to the infinite spatial limit—only those parameters relevant to phase transitions will remain in the renormalized Lagrangian. One of the most qualitatively interesting features of successively averaging out short distance (and therefore high energy) degrees of freedom is that, no matter how complicated the initial field dynamics are (encoded as a Lagrangian), only the renormalizable terms will contribute to the low energy dynamics of the theory. This implies that a very broad class of higher energy Lagrangians can “reduce” to the relevant dynamics at lower energy scales.

The success of RG methods in CMP lead to their quick application in QFTs (Wilson 1983)⁶, since the relevant formalism is shared between the two disciplines. If we choose

⁶Wilson even forms the QFT/statistical mechanics analogy explicitly, though the source analog in that

to endow the RG methods with similar physical significance in QFT, then we can interpret the high energy cutoffs required as marking the domain at which we expect new physics to occur. The problem is that, because RG flows tell us that our low-energy (effective) QFTs are largely insensitive to the dynamical details at higher energies, they provide little insight our guidance into the high energy physics. Though the path to the successor theory isnt apparent given our current QFTs, the up side is that our best QFTs are protected from the details of our ignorance of high energy dynamics.

Where Wallace might be read to err is in the jump from believing that cutoffs have physical relevance in QFTs to believing that cutoffs *are physical*:

“This, in essence, is how modern particle physics deals with the renormalization problem: it is taken to presage an ultimate failure of quantum field theory at some short lengthscale, and once the bare existence of that failure is appreciated, the whole of renormalization theory becomes unproblematic, and indeed predictively powerful in its own right” (Wallace 2011, p. 119).⁷

The difference is subtle. Cutoffs can be *physically relevant* in that they signal the breakdown of the *particular* theory or model beyond a certain energy scale, but whether cutoffs themselves *are physical* depends on the precise nature of the breakdown. If the

case is a classical Ising model (Wilson and Kogut 1974). Fraser (2016) has provided an in-depth analysis of the elements of the analogies between QFT and the Ising model, as well as the process of describing RG flow.

⁷Or at least this is a jump he is sometimes guilty of. In other places he is more careful to elaborate on this view, and it appears that he at least appreciates the fact that field theoretic methods may not break down at all (Wallace 2006, pp. 43-4). As mentioned in the introduction, this paper is not a critique of Wallace’s view explicitly, but of the misleading way of framing AQFT and CQFT as rivals based on their differing treatments of the arbitrarily small; for this reason I aim to clarify the mistakes in a “naive” reading of Wallace.

breakdown can be remedied by adding new terms in the Lagrangian—effectively changing the particular theory, but retaining the field theoretic framework—then the cutoffs signal new physics, but are not themselves physical. If the breakdown is due to the inapplicability of field theoretic methodology beyond that scale, then the cutoffs are themselves physical.⁸ Even if one takes the cutoffs to have physical significance, cutoffs needn't *be physical* in this stronger sense.

One possible reason for thinking that cutoffs are physical is based off of reading too much into the analogy with CMP. We know that field theoretic methods are approximations in bulk matter systems—the atomic theory implies that macroscopic matter is composed of discrete components. The analogy between QFT and CMP is based on the use of the same field theoretic formalism in both disciplines, not on a well-grounded physical similarity.⁹ Cutoffs are physical in CMP field theory because field theoretic methods have been introduced as an approximation. Given that discrete quantum mechanics of 10^{23} particles is intractable, we sacrifice (a surprisingly small amount of) precision in order to apply the more soluble methods developed in QFT. But the fact that cutoffs signal the breakdown of field approximations in CMP does not imply that the same is true in QFT. The reasons we treat cutoffs as physical in CMP are absent in QFT; there is no empirically successful theory that claims QFT breaks down due to an underlying discreteness of physics near cutoff scales. Speculative physics may posit some underlying structure for which quantum fields are merely an approximation,

⁸Presumably, the failure of field theoretic methodology in general would require some physical granularity at high energies. This is what I mean by the cutoff being physical and is in direct analogy with the case of non-relativistic QFT in CMP.

⁹Fraser (2016) and Fraser and Koberinski (2016) provide two concrete examples of fruitful formal analogies between QFT and CMP. In the former case, it is the RG flow that is formally analogous, while the latter deals with the formal similarities between spontaneous symmetry breaking within the two theories.

but until any of these theories make successful empirical predictions their significance for interpreting QFTs must be limited.

2.4 Why physical cutoffs are also a problem for CQFT

Even though, as I have argued, there is currently no physically motivated reason for supposing cutoffs to be physical, it may be the case that we find such a reason in the future. Perhaps we will need radically different methods from those of field theory to describe physics beyond the Standard Model. There is no shortage of candidates that claim to radically alter our picture of the world—from 11-dimensional string theory to discrete spacetime to the emergent spacetime of loop quantum gravity. Though experimental support for any of these speculative theories would mean that the axioms of any AQFT must be at best only approximations, this does not mean that CQFT would escape unscathed. Any observed violation of Lorentz invariance would signal bad news for both AQFT and CQFT, and the extent to which we choose to reject or salvage the former, we should do the same for the latter.

Though its importance is not encoded in a set of axioms, Poincaré invariance is of central importance to the physical content of CQFT. In constructing QFTs, one starts by writing down a classical Lagrangian to encode the physical content of the theory. The two major constraints on the form of candidate Lagrangians are renormalizability (dealt with above) and Poincaré invariance. Since the Lagrangian is a scalar, it must remain strictly invariant under the action of the Poincaré group on its component fields. All of the fundamental forces—as described by the Standard Model—are encoded in Lagrangians obeying strict Poincaré invariance. If anything qualifies as physically relevant to CQFT, the Lagrangian certainly does; it is the starting point for building a

QFT, and determines the types of fields, their masses, and the particulars of their interactions. A violation of Poincaré invariance at a more fundamental level—be it in a particular physical process or in the structure of some new spacetime picture—undercuts to the same extent the physical significance of *any and all* theories that depend on Poincaré invariance for their formulation. Thus, despite the lack of rigid and precise axioms demanding Poincaré invariance, the physical content of CQFT stands or falls with AQFT.¹⁰

Once again, the major difference between AQFT and CQFT lies in the formalism. Though the *physical* content of CQFT is built upon Poincaré invariance¹¹, the formalism is indifferent to the constraints placed upon the Lagrangian. The success of field theoretic methods in CMP is evidence of the flexibility of the formalism; in CMP the Galilean group is taken as the appropriate symmetry group, given the low energies dealt with. In contrast, the formalisms of various AQFTs are constructed around the axioms. Any theorems that rely on exact Poincaré invariance will only hold in the real world if nature is Poincaré invariant.¹² The greater precision of the formalism in AQFT makes it more rigid in this regard.

If violations of Poincaré invariance are problematic for all variants of QFT, should investigators into the foundations of QFT fret if such violations are experimentally

¹⁰CQFT *methods* could still be useful, but the theoretical framework of CQFT—as encoded in the Standard Model—depends on Poincaré invariance.

¹¹Depending on how one views Poincaré invariance, this may seem odd. The specific transformation properties of scalars, vectors, and tensors under the Poincaré group are undoubtedly formal properties of the particular field representations. However, the physical symmetries represented in this way have a physical basis (e.g., rotation invariance implies that the physical system can be modelled the same way when rotated).

¹²Though it isn't always possible, proofs of the form “If Minkowski spacetime then *x*” are strengthened and made more robust by also showing “If *approximately* Minkowski spacetime then *approximately x*.” Given that our best current theories lead us to believe that spacetime is only locally Minkowski, these are the results for which we can have a high degree of confidence in their robustness.

confirmed? No; the experimental success of QFT implies that the world is at least *approximately* Poincaré invariant, and any evidence revealing the limits of that approximation has no bearing on the theory itself. We have good reason to believe that the QFTs in the Standard Model are not the final story: General Relativity implies that strong gravitational effects distort spacetime, and that our spacetime is only ever Minkowski in small patches where gravity is negligible. Though this approximation seems to hold for experiments at the LHC, if we want a theory that gets spacetime symmetries *exactly* correct, QFTs relying on Poincaré invariance will not do the trick. Rather than abandoning foundations of QFT for being approximate at best, investigation should proceed given that QFTs are highly successful within the energy domain currently testable. To this extent, we are justified in viewing the world as approximately described by QFTs, and should content ourselves with investigating an incomplete (though highly accurate) picture of nature. Whether we are dealing with a formalism that encodes Poincaré invariance into its axiomatic framework, or a formalism in which Poincaré invariance has been used indirectly to construct empirically successful theories, we should not take violations of Poincaré invariance as signalling the failure of either approach. Any robust results obtained within either formalism will still hold approximately, and should be equally subject to foundational analysis.

3 Conclusions

I have tried to show that cutoffs do not provide physical grounds for separating AQFT and CQFT as rival research programs. First, RG methods can be incorporated into AQFT without major issue, and cutoffs can be introduced as well—though explicit

cutoffs provide a more pressing conceptual revision to AQFT. Second, we needn't take AQFT to be an exact description of the world. In the same way that classical fluid dynamics is compatible with classical point mechanics, AQFT defined to arbitrary precision can be compatible with a CQFT that requires cutoffs. The appropriate lesson is that we should take AQFT to be approximately true in sufficiently low energy domains. Finally, even if cutoffs are of physical significance, they don't require a breakdown of continuum methods in general. This idea stems from pushing an analogy with CMP, which appears to be unjustified.

Though the Fraser-Wallace debate has spawned increased investigations into the foundations of QFT, it has set the boundaries of the debate in such a way as to create a false dichotomy: one is forced to choose whether to immerse oneself in the AQFT or CQFT formalisms. When we discard the false dichotomy and recognize AQFT as complementary to CQFT, we open the door to the synthesis of axiomatic methods with Lagrangian QFT. In this way the general features of QFTs can be investigated rigorously in AQFT, and we can be confident that—insofar as the axioms of AQFT capture the physical assumptions of CQFT—the results carry over to CQFT.

Though it is true that there do not yet exist AQFT models that incorporate interactions in four-dimensional spacetime, the successes of AQFT have been compatible with CQFT. Free field theories and ϕ_2^4 interaction theories constructed in AQFT give predictions in agreement with comparable CQFTs. Insofar as AQFT is a successful formalism, its results should be thought of as complementary to those of CQFT: one uses the same physical principles to construct differing formalisms.

In essence, I advocate for a position similar to Wallace's earlier view (though note that in this passage he refers only to specific results of AQFT, such as the spin-statistics

theorem):

the foundational results which have emerged from AQFT have been of considerable importance in understanding QFT and in general they apply also to Lagrangian QFTs. This paper should be read as complementary to, rather than in competition with, these results (2006, p. 35).

The particular choice of formalism will depend on the scope of the foundational investigation. If the goal is to prove general results applicable to any relativistic QFT, then AQFT is the appropriate formalism; if the goal is to determine the consequences of specific physical interactions, then CQFT should be used.

References

- Buchholz, Detlev and Rainer Verch (1995). “Scaling algebras and renormalization group in algebraic quantum field theory”. In: *Reviews in Mathematical Physics* 7.8, pp. 1195–1239.
- Fraser, Doreen (2009). “Quantum field theory: Underdetermination, inconsistency and idealization”. In: *Philosophy of Science* 76, pp. 536–567.
- (2011). “How to take particle physics seriously: A further defence of axiomatic quantum field theory”. In: *Studies in History and Philosophy of Modern Physics* 42, pp. 126–135.
- (2016). “The development of renormalization group methods for particle physics: Formal analogies between classical statistical mechanics and quantum field theory”. Forthcoming in *The British Journal for the Philosophy of Science*.
- Fraser, Doreen and Adam Koberinski (2016). “The Higgs mechanism and superconductivity: A case study of formal analogies”. Forthcoming in *Studies in the History and Philosophy of Modern Physics*.
- Haag, Rudolf and Daniel Kastler (1964). “An algebraic approach to quantum field theory”. In: *Journal of Mathematical Physics* 5.7, pp. 848–861.
- Halvorson, Hans and Michael Müger (2007). “Algebraic quantum field theory”. In: *Handbook of the Philosophy of Physics, Part A*. Ed. by Jeremy Butterfield and John Earman. Elsevier.
- Ruetsche, Laura (2011). *Interpreting quantum theories*. Oxford University Press.
- Wallace, David (2006). “In defence of naiveté: The conceptual status of Lagrangian quantum field theory”. In: *Synthese* 151, pp. 33–80.

- Wallace, David (2011). “Taking particle physics seriously: A critique of the algebraic approach to quantum field theory”. In: *Studies in History and Philosophy of Modern Physics* 42, pp. 116–125.
- Wilson, Kenneth (1983). “The renormalization group and critical phenomena”. In: *Reviews of Modern Physics* 55.3, pp. 583–600.
- Wilson, Kenneth and John Kogut (1974). “The renormalization group and the ϵ expansion”. In: *Physics Reports* 12.2, pp. 77–199.

**On Epistemically Detrimental Dissent:
Contingent Enabling Factors v. Stable Difference-Makers.**

Soazig Le Bihan and Iheanyi Amadi

Abstract.

The aim of this paper is to critically build on Justin Biddle and Anna Leuschner's characterization (2015) of epistemologically detrimental dissent (EDD) in the context of science. We argue that the presence of non-epistemic agendas and severe non-epistemic consequences are neither necessary nor sufficient conditions for EDD to obtain. We clarify their role by arguing that they are contingent enabling factors, not stable difference-makers, in the production of EDD. We maintain that two stable difference-makers are core to the production of EDD: production of skewed science and effective public dissemination.

Introduction.

The aim of this paper is to critically build on Justin Biddle and Anna Leuschner's characterization of epistemologically detrimental dissent (EDD) in the context of science (2015). We follow their lead in taking 'dissent' to be a particular kind of criticism, i.e. the act of objecting to a widely held conclusion. When done properly, dissent is welcome within scientific practice. As Helen Longino has clearly established, "scientific knowledge is produced collectively through the clashing and

meshing of a variety of points of view (1990, 69). Criticism, when done properly, is integral to the collective advancement of science.¹ Dissent, when an instance of proper criticism, is thus epistemically valuable in the context of science.

Now there are some instances of dissent that come out as epistemically detrimental. That is to say, some instances of dissent seem to impede, not promote, the collective advancement of science. Many examples come to mind, that have been well described in the recent literature (Oreskes and Conway 2010, Biddle and Leushner 2015, Harker 2015). Roughly speaking, EDD is about manufacturing controversy in a particular scientific field. The typical story goes something like the following. The research involved has some severe non-epistemic consequences in terms of, on one side, industry profit, and, on the other side, public welfare; large amounts of money are invested by industry-related groups to (1) produce some skewed research, (2) largely publicize the results through the media, (3) produce an atmosphere of confusion and doubt within the public, (4) launch some campaign against the lead scientists of the field in the media and political world (often through personal attacks and threats); this results in an atmosphere in which the scientists subjectively feel a lot of pressure and discomfort, and also objectively waste precious time and limited resources to address the well-publicized skewed research. At this point, the collective advancement of science is clearly impeded. We have an instance of EDD.

¹ Longino (1990) offers an account of some of the various kinds of epistemically beneficial criticism within science.

The aim of this paper is to properly distinguish, in that story, between (1) contingent enabling factors, and (2) stable difference-makers, in the production of EDD. Our most contentious claim is that the intrusion of non-epistemic agendas and presence of severe non-epistemic risks are contingent enabling factors, not stable difference-makers for EDD. We maintain that two stable difference-makers are core to the production of EDD: production of skewed science and effective public dissemination.

In Section 1, we offer what we take to be the most straightforward argument for the claim that intrusion of non-epistemic agendas is not sufficient in the production of EDD: it may lead to EDD only if it leads to skewed science. In Section 2 we argue that it is not necessary either. Section 3 is devoted to a clarification of the role of intrusion of non-epistemic agendas in EDD on the basis of a distinction between contingent enabling factors and stable difference-makers. Section 4 investigates the consequences of our analysis for the Inductive Risk Account of EDD proposed by Biddle and Leuschner (2015).

Section 1. Non-epistemic agendas: not sufficient for EDD

That intrusion of non-epistemic agendas is not sufficient to the production of EDD has been discussed by Wilholt (2009), and Biddle and Leuschner (2015). Roughly, the point is simply that, unless intrusion of background non-epistemic agendas is such that the work produced *fails to satisfy some of the conventional standards for proper science*, there is no problem. We offer here what we take to be the most straightforward argument for this point.

As the community of philosophers of science have recently come to recognize, intrusion of non-epistemic values in scientific practice is quite common (Douglas 2009). Now obviously, that does not necessarily result in skewed science. If a scientist defends a conclusion C on the basis of evidence E, the fact that some background non-epistemic values enters in her reasoning does not matter if (1) she can publicly produce a reasoning in defense of C, and if (2) that reasoning can be assessed as adequate scientific reasoning by her peers, including peers who do not share the same background non-epistemic values. If these two conditions are met, then the conventional standards for proper science are met, and we do not have a case of skewed science. Now if proper scientific work was produced, there is no a priori reason to think that her work cannot partake in the collective advancement of scientific knowledge. It might do so at various degrees, but that will depend on its heuristic value, which is a priori unrelated to whether or not there was intrusion of non-epistemic values.

Let us push this line of argument a little further. It is important here to underline the fact that the reasoning rendered public by the scientist might not be the actual reasoning through which she came to accept either E or its relevance with regard to C. From a subjective point of view, for example, she might well have had accepted C well before she produced E and the reasoning defending the relevance of E as supporting C. She might well have accepted C for non-epistemic, value-laden, reasons. However, such considerations over the subjective state of scientists do not matter. The collective assessment of scientific research is not in the business of mind reading. No matter what kind of reasoning (or non-reasoning) actually

brought a scientist to believe C, the relevant question is whether she is capable of producing a reasoning in defense of E and its relevance with regard to C that can be publicly, and positively, assessed by the experts in her field. To put it bluntly: the most biased and ill-intentioned scientists are a priori capable of producing good scientific work.²

This line of argument applies to the production of dissenting views. Dissenting claims proposed by scientists motivated by non-epistemic agendas do not necessarily lead to skewed science and hence to of EDD. If a reasoning can be publicly produced, and if the members of the scientific community, including members of that community who do not share the same values as the dissenting views' proponents, assess that reasoning as scientifically adequate, then we do not have an instance of skewed dissent. As an instance of work that satisfies the agreed-upon standards of proper scientific practice, the dissenting view could well participate in the advancement of scientific knowledge. It could do so at various degrees, depending on how important the dissenting views are, but that would not depend on whether or not the dissenting views are the product of scientists with non-epistemic agendas. Considerations about the subjective intentions, or background beliefs, of the scientists are irrelevant, unless one can show that skewed science was produced.

² This is not denying the actuality of implicit bias. By definition, implicit bias is still bias. As such, it can be recognized by the scientific community for what it is. What is implicit about it is that the biased author (and possibly some of her peers as well) is not even realizing her own bias.

Section 2 Non-epistemic agendas: not necessary for EDD

At this point, we have shown that intrusion of non-epistemic agendas do not necessarily result in the production of EDD. Note that EDD does not require intrusion of non-epistemic agendas either. What would it take to have a case of EDD without any intrusion of non-epistemic agendas? We know that EDD is about manufacturing controversy within a scientific field. First, the controversy is “manufactured”, not genuine, because the dissenting view is not based on proper science; it violates some of the commonly accepted standards for proper scientific practice; it is an instance of skewed science. Now skewed science can come to be in many ways. It does not have to result from the intrusion of non-epistemic agendas. One can imagine the case of a scientist, say Jack, who is genuinely interested in partaking in the collective advancement of scientific knowledge, but is also a poor scientist. One can imagine that Jack is very wealthy, and thus has both the time and financial resources to pursue his research, and produce a large amount of work challenging the commonly held views in a given scientific field. Jack, albeit misguided in many ways, could conceivably do all of this with the “purest” goal in mind.

Now one immediately sees that the production of bad science is not enough to produce EDD. Jack’s research is likely to be simply ignored by the scientific community. So what would it take to “manufacture” a controversy on the basis of Jack’s research? The answer seems rather straightforward: Jack’s research needs to be effectively disseminated, so that scientists feel pressured to respond to Jack’s

challenges. The standard avenues for dissemination of scientific research, i.e. peer-reviewed publication, however, are not likely to be an option for Jack, since his work is widely recognized by the community as being of poor scientific quality. He must then bypass these avenues, and manage to effectively disseminate his research among the public. Mass media would be a likely option for this. This in turn forces scientists in the field to waste time and resources to address Jack's research. Hence a case of EDD, with the purest epistemic goal at its source.

The case above might seem far-fetched. One objection could be that, unless some non-epistemic values were at stake, it is unlikely that the media and the public would get interested in Jack's research, and Jack would fail to be able to manufacture the controversy. It might be unlikely, but it is surely conceivable. If Jack's public dissemination machinery is effective enough, (mis-) understandings over the state of research in the field of concern could well have serious repercussions on public funding. Jack could well have a very strong network of communication – he could well be the owner of a very large cable and press network. Repeated reporting on public funding of supposedly controversial science could well spur outrage in the public. "Debates" on mass media would ensue. As soon as the scientists would engage in that conversation, Jack's claims would gain in credibility.³ At the end, Jack's campaign could well be so effective that scientists

³ This is a point that Hannah Arendt made clear in her insightful analysis of controversy- and doubt-manufacturing in a completely different context, i.e. the (non-)issue of the reality of the Holocaust during WWII (1966/2010).

would indeed be forced to repeatedly address his research to defend their own. So, intrusion of non-epistemic agendas is not necessary to the production of EDD.

Section 3. Stable Difference-Makers v. Contingent Enabling Factor

From the discussion above, we conclude that intrusion of non-epistemic agendas is neither necessary nor sufficient for the production of EDD. Such a conclusion might strike many as unsatisfactory, however. Isn't it the case that intrusion of non-epistemic agendas was an important factor in the production of the common cases of EDD that we have witnessed over the last 50 years? Some may even want to claim that, as a matter of fact, in all of the cases we know of in recent history, no EDD would have occurred if it were not for the intrusion of non-epistemic agendas. This is an important intuition, and arguably, any satisfactory account of EDD ought to make sense of it. Fortunately, we believe there is a way to do so, that is, by appealing to the distinction between contingent enabling factors and stable difference-makers as discussed by Thomson (2003) and Woodward (2010). Thomson (2003) makes the point (contra many theories of causation) that just because 'E would not have happened without C', it does not follow that 'C has caused E'. She argues that the proposition 'E would not have happened without C' only entails that 'C was physically necessary for E'. Consider her example. John built a bridge over the Rapid River. The Rapid River is notoriously wild, and only John, a master-builder, could have done it. From the bridge being built, it ensues that Smith crosses the river. Now John's building the bridge was physically necessary to Smith's crossing the Rapid River, but most would agree that it is misguided to take it

as a cause for it. John's building the bridge, even if "physically necessary" in the whole process, remains largely irrelevant to Smith's crossing the river. It belongs to the background conditions, or environmental conditions, that make Smith's crossing possible, without causing it in any genuine sense of causation. In Thomson's vocabulary, it is only an enabling factor.

Woodward (2010) is interested in analyzing a similar distinction between the core difference-makers and the background conditions. His analysis is useful to flesh out some of the characteristics of enabling factors à la Thomson.⁴ One of intuitions Woodward is trying to capture is that some causal relationships are robust, i.e. insensitive to environmental change, while others are contingent on the presence of a specific environment. To do so, he articulates the notions of "stability".⁵ A causal relationship, according to Woodward, is stable if and only if it holds over a wide range of background conditions. Some examples might be useful at this point.

⁴ Note that we do not claim (and neither does Woodward) to have unveiled the set of necessary and sufficient conditions for factors to qualify as enabling factors by contrast to stable difference-makers. We will only claim that being enabling factors are typically unstable, and hence, that lack of stability serves as a good indicator for a factor to be only enabling, not causing.

⁵ Two other notions are articulated in the article. The notion of proportionality serves to address the issue of the proper levels of explanation. The notion of specificity serves to address the issue of coarse v. fine-grain causal influence.

A paradigmatic example of an unstable relation would be the following.⁶ “Star” professor P writes a letter of recommendation for Jane, thanks to which Jane gets a job at university U. She would not have gotten the job without it. Jane meets Joe at U, they get married, and have children. Challenged by the difficulties of coupling an academic career with quality parenting, Jane goes into depression. Now consider the following claim: ‘P’s writing a letter for Jane caused Jane’s depression’. Given the story that is given, there is a sense in which P’s writing a letter for Jane enabled Jane’s suffering from depression, but there is also a strong sense in which it is misguided to take it as a cause for it. The reason is that the relation between P’s writing the letter and Jane’s suffering from the disease would cease to hold under many small, contingent, changes in the background conditions for the story (Jane and Joe could not have met, they could have decided to not have children, U could have had a very progressive parental leave policy, etc.). The causal relationship between the letter and the depression is thus highly unstable because it holds only in a very specific environment.

Now contrast this with a paradigmatic example of a stable relation. I turn on the heat under my closed pressure cooker (with some water in it). The pressure goes up and the valve shuts down. Clearly, heating up the pressure cooker is a stable cause of the pressure valve to shut down. Many of the most stable causal relations are backed up by what the kind of generalizations that we take to be the laws of physics, or chemistry. These generalizations hold over a wide range of background conditions.

⁶ This example is inspired by Woodward (2010) himself inspired by Lewis (1986).

There are obviously various degrees of stability in between these two extreme cases. Stability is not an all or nothing affair. It might also be difficult to figure out which causal relationships are more or less stable. That said, it could also be worth the effort looking into it, because, how stable a factor is could be a measure of how well we can target change by targeting that factor in a given situation. As Woodward explains (2010, 315): “other things being equal, causal relationships that are more stable are likely to be more useful for many purposes associated with manipulation and control than less stable relationships.” Applied to our case, if ultimately we hope to be able to alter the manufacturing of controversy and EDD, it could turn out to be very useful to clarify the causal landscape behind EDD by distinguishing between the contingent enabling factors and the more stable difference-makers.

Thomson’s and Woodward’s analyses are clearly related. Thomson’s bridge example is a clear case of a very unstable causal relationship: it holds only under very specific background conditions (The Rapid River could have been gently, Smith could have decided not to cross the bridge, etc.) Some unstable causal relationships as discussed by Woodward are so at least partially because they are relationships of contingent “physical necessity” à la Thomson. So, a causal factor may be highly unstable, despite being ‘necessary’ to the causal process, if its influence on the process is highly contingent on a specific environment. No matter how “necessary”

in that sense a factor F is, F being unstable points F being an enabling factor, not a stable difference-maker.⁷

The discussion above allows us to bring home two important points. First, it allows us to identify two stable difference-makers for the production of EDD: the production of skewed scientific research and its effective public dissemination. That the combination of these two factors produces an instance EDD holds over a wide range of conditions. What changes in background conditions would make that causal relation fail? First, one could think of a world in which scientists could ignore even well-advertised skewed science. For example, that could possibly be the case in a world in which production of scientific research would not depend on getting public funding, or in a world in which the public is generally knowledgeable about (the philosophy of) science, and hence, is able to recognize that the well-

⁷ Two points of clarification are in order. First, Woodward convincingly argues that the extent to which a cause is stable is related, but not equivalent to, its distal/proximate character vis à vis the effect. Second, Woodward also argues that stability is not dependent on the level of explanation: degrees of stability are not necessarily to how “reductive” the explanation is. So, our distinction between contingent enabling factors and stable difference-makers is not trivial in the sense that the most stable difference-makers would always be the most proximate causes described at the level of fundamental particles.

advertised science is skewed. Arguably, these do not qualify as small changes in the background conditions for scientific practice.⁸

The second point is a clarification of the role played by the intrusion of non-epistemic agendas in the production of EDD. Intrusion of non-epistemic agendas is not a stable difference-maker for the production of EDD. This is because there is a large range of conditions under which intrusion of non-epistemic agendas do not result in EDD. These include the conditions for all the cases in which intrusion of non-epistemic agendas do not result in skewed science. If we take seriously recent work on science and value, intrusion of non-epistemic values is actually the rule, not the exception within the practice of science (Douglas 2009, Intemann 2001, 2015, and references therein). Note that, if our take on Thomson's and Woodward's analyses is correct, then the claim that intrusion of non-epistemic agendas is not a stable difference-maker but only a contingent enabling factor is consistent with the fact that it has been "physically necessary" in many of the well-known instances of EDD. One can consistently say that, while not a stable difference-maker, it has been an important enabling factor for the production of well-publicized skewed science. Intrusion of non-epistemic agendas has been necessary for some groups to develop an *interest in funding* the production and public dissemination of skewed research.

⁸ There is also a possibility that some cases of EDD could come out of seemingly proper science "distracting" the public from the most widely held views within the scientific community. We believe that even in these cases, dissenting views do not entail EDD unless there is violation of some conventional standards for proper science. This interesting issue belongs to another paper.

That said it is important to distinguish between factors that are characterized by this kind of ‘necessity’ (the bridge or letter kind of necessity) and factors that are true stable difference-makers. It is all the more important that, if one of our goals is to alter the production of EDD, then our analysis suggests that intrusion of non-epistemic agendas is not the proper target. Once again, non-epistemic values are the common rule within the practice of science. A more efficient approach in the prevention of EDD would be to understand the various ways skewed science may be produced. This includes the important discussion on the distinction between legitimate and illegitimate use of non-epistemic values in scientific practice (Hicks 2014, Intemann 2015). This in turn includes an investigation of the mechanisms by which intrusion of non-epistemic values does result in skewed science. Implicit bias might one of these mechanisms. Inductive risk bias, as we shall explain in the next section, is another one. Before we turn to this point, let us take stock.

We have clarified the causal landscape for the production of EDD. We have identified two stable difference-makers – production of skewed science and its effective public dissemination; and we have characterized the important role of intrusion of non-epistemic agendas within science as contingent enabling factors for the production and dissemination of skewed research, hence for EDD.

Section 4. Consequences for the Inductive Risk Account of EDD

Biddle and Leuschner have articulated what they call the “inductive risk account” of EDD (2015). According to this account, the following set of conditions are jointly sufficient for the production of EDD (2015, 273):

Dissent from a hypothesis H is epistemically detrimental if each of the following obtains:

- (1) The non-epistemic consequences of wrongly rejecting H are likely to be severe*
- (2) The dissenting research that constitutes the objection violates established conventional standards.*
- (3) The dissenting research involves intolerance for producer risks at the expense of public risks.*
- (4) Producer risks and public risks fall largely upon different parties.*

Biddle and Leushner admit that these conditions are not necessarily related to the production of EDD (275):

“We are not arguing that, in all possible worlds, research that meets the conditions of the inductive risk account inhibits the progress of science. It is possible, for example, to organize science and to regulate industry in such a way that dissent that meets these conditions is not widely disseminated, does not acquire political authority, and is not used to attack mainstream scientists. But this is not the way in which science and society are currently organized. Dissent that meets the conditions of the inductive risk account is, given current societal arrangements, likely to inhibit knowledge production, particularly because of the success of political, economic, and ideological interests in structuring the dissemination of research.”

We think that the framework used in Section 3 can help clarify the causal landscape for the production of EDD offered in the Inductive Risk Account. Our contention is that Biddle and Leuschner, by focusing on inductive risk, have identified a

particular, important, but still contingent, enabling factor, but have failed to clearly distinguish the proper core of stable difference-makers, for the production of EDD.

Let us make that point in more details.

The four conditions above can be seen as dividing into three groups. Condition (2) identifies one of the stable difference-makers – production of skewed science. Conditions (1) and (4) together specify some particular enabling conditions for the formation of non-epistemic agendas – the presence of severe and opposing non-epistemic consequences (SONEC). Condition (3) identifies a mechanism by which intrusion of SONEC-related non-epistemic agendas may enable the production of skewed science. In other words, the inductive risk account of EDD identifies an important series of enabling causes leading to one of the two stable difference-makers we have identified in Section 1-3, i.e. production of skewed science. That series of cause is something like this: from the presence of SONEC to biased inductive risk reasoning, and to skewed science. This is an important contribution to the understanding of EDD precisely because it not only identifies some particular enabling factors (the presence of SONEC) for the formation of epistemic agendas, but also a mechanism by which intrusion of SONEC-related non-epistemic agendas may enable the production of skewed science (via inductive risk bias). Now it is also important to clarify the causal landscape and recognize that fulfillment of Condition (2) is the stable difference-maker which fulfillment of Conditions (1), (4), and then (3) enable as a matter of contingent fact. Biddle and Leuschner seem to have missed that useful distinction.

If our analysis in Section 3 is correct, they also have failed to include the second stable difference-maker for EDD, i.e. effective public dissemination. As they admit in the paper (see quote above), the presence of SONEC obviously does not imply that effective public dissemination will ensue. Conversely, as Jack's case shows, effective public dissemination could well be obtained without the presence of SONEC. How (un-)likely this is obviously is an empirical question. No matter how unlikely, however, it is important for our understanding of EDD to mention effective public dissemination as a core stable difference-maker. The inductive risk account fails to do so. Let us underscore, however, that Biddle and Leuschner once again have identified an important mechanism by which presence of SONEC enables effective public dissemination and the manufacturing of controversy: the presence of SONEC not only enables the production of skewed science, but also the establishment of "sophisticated, private-funded network for disseminating [dissenting] results" (2015, 275).

This brings us to our conclusion on the Inductive Risk Account: Biddle and Leuschner have successfully identified an important contingent enabling factor for EDD, i.e. the presence and influence of SONEC. That said, they have failed to distinguish between the different roles that enabling factors and stable difference-makers play in the production of EDD. We hope to have clarified the situation.

Conclusion

Well-known cases of EDD seem to have in common various forms of intrusion of non-epistemic, often SONEC-related, agendas within the science. We have argued

that such intrusion is not core to the production of EDD: neither necessary nor sufficient, it is also not a stable difference-maker. We have clarified its causal role: intrusion of non-epistemic agendas is a contingent enabling factor. Reduced to its core, EDD is just well-advertised bad science. Because it is well advertised, it has an impact on the collective building of scientific knowledge. Because it is bad science, it does not advance that endeavor, but any case negatively impacts it instead.

To make the distinction between contingent enabling factors and stable difference-makers is important for at least three reasons. First, it is important to clarify the causal landscape that leads to the production of EDD, as it simply increases our understanding of EDD. Second, it might suggest more efficient avenues for targeting change. Finally, it is crucial to make room for the intrusion of non-epistemic values within the science without it being epistemologically detrimental. As the community of philosophers of science comes to recognize that such intrusion is the rule rather than the exception, one must leave conceptual room for a distinction between “legitimate” and “illegitimate” role for non-epistemic values within science (Hick 2014, Intemann 2015).

Bibliography

Arendt, Hannah. 1967/2010. “Truth and Politics.” In José Medina and David Wood (eds). *Truth. Engagements Across Philosophical Traditions*. Blackwell: 295-314.

Biddle, Justin B. and Anna Leuschner. 2015. "Climate Skepticism and the Manufacture of Doubt: Can Dissent in Science be Epistemically Detrimental?" *European Journal for Philosophy of Science* 5 (3): 261-278.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Harker, David. 2015. *Creating Scientific Controversies: Uncertainty and Bias in Science and Society*. Cambridge University Press.

Hicks, Daniel J. 2014. "A New Direction for Science and Values." *Synthese* 191 (14): 3271-3295.

Intemann, Kristen. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5 (2): 217-232.

———. 2001. "Science and Values: Are Value Judgments always Irrelevant to the Justification of Scientific Claims?" *Philosophy of Science*: S518.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Lewis, David. 1986. "Postscript c to 'causation': (insensitive causation)" in: *Philosophical papers*, vol 2. Oxford University Press, Oxford: 184–188

Oreskes, Naomi and Erik M. Conway. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Publishing USA.

Thomson, Judith Jarvis. 2003. "Causation: Omissions." *Philosophy and Phenomenological Research* 66 (1): 81-103.

Wilholt, Torsten. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science Part A* 40 (1): 92-101.

Woodward, James. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology & Philosophy* 25 (3): 287-318.

Literal vs. careful interpretations of scientific theories: the vacuum approach to the problem of motion in general relativity

Dennis Lehmkuhl
Einstein Papers Project and HSS Division,
California Institute of Technology
Email: lehmkuhl@caltech.edu

Forthcoming in *Philosophy of Science* (PSA 2016 Supplement)
Version: September 26, 2016

Abstract

The problem of motion in general relativity is about how exactly the gravitational field equations, the Einstein equations, are related to the equations of motion of material bodies subject to gravitational fields. This paper compares two approaches to derive the geodesic motion of (test) matter from the field equations: ‘the T approach’ and ‘the vacuum approach’. The latter approach has been dismissed by philosophers of physics because it apparently represents material bodies by singularities. I shall argue that a careful interpretation of the approach shows that it does not depend on introducing singularities at all, and that it holds at least as much promise as the T approach. I conclude with some general lessons about careful vs. literal interpretations of scientific theories.

Contents

1	Introduction	2
2	A critical comparison	5
3	The vacuum approach	8
3.1	Two ways of looking at Einstein’s model of the Sun-Mercury system	8
3.2	The Einstein-Grommer vacuum approach to the problem of motion	9

4 Interpreting Einstein-Grommer	11
5 Conclusion	15

1 Introduction

It is a bit of an irony that one of the most widely embraced definitions of what it means to be a scientific realist is due to the arch-anti-realist Bas van Fraassen. His definition starts by stating that “Science aims to give us, in its theories, a literally true story of what the world is like”.¹ And indeed, scientific realists often see themselves as committed to ‘taking scientific theories at face value’: if the best theories of particle physics say that quarks exist, then we should believe that they exist; if general relativity tells us that gravity is really just an aspect of spacetime structure, then we should believe it; if quantum mechanics tells us that the world is at its core non-deterministic, then we should believe that too.

The problem is that scientific theories, or at least the theories of modern physics, are not that straightforward with us. They may seem so at first, but if you listen to the details of their respective stories, if you take your time to look under the surface, what exactly we should take them to tell us about the world is far from clear. Murray Gell-Mann, the inventor of the concept of quarks, for a long time did not think that quarks should be interpreted as literally existing; neither did Richard Feynman. Albert Einstein passionately resisted the interpretation of general relativity that says that the gravitational force field of Newtonian theory is ontologically reduced to the geometry of spacetime in general relativity. And of course, there is a long-standing battle in foundations of physics about whether quantum mechanics really does tell us that the world is non-deterministic.²

In this paper I shall introduce a new case study that provides further evidence for the position that, whether you are a realist or not, the *literal interpretation* of a scientific theory, especially in physics, can be rather misleading. I will argue that what we should aim for is a *careful interpretation*;

¹Van Fraassen [1980], p.8.

²For a discussion of different interpretations of the quark concept see Pickering [1999], for Einstein’s opposition to interpreting general relativity as a geometrization of gravity see Lehmkuhl [2014], and for debate on whether quantum mechanics is really indeterministic see e.g. Saunders et al. [2010].

an interpretation of the theory or model or formalism that engages with its details, both with the details of its mathematical structure and with how it is applied to the natural world. Philosophy of science must be willing to look under the hood.

The case study I want to look at is the so-called problem of motion in the general theory of relativity (GR). It asks about the precise relationship between the two sets of equations that are at the very heart of GR. On the one hand there are the Einstein field equations, which give us the dynamics of the gravitational potential (the metric tensor) $g_{\mu\nu}$:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu} = \kappa_E T_{\mu\nu} \quad . \quad (1)$$

On the other hand, we have the geodesic equation that determines which paths through spacetime are geodesics of the connection $\Gamma^\nu_{\mu\sigma}$ compatible with the metric $g_{\mu\nu}$:

$$\frac{d^2 x_\tau}{ds^2} + \Gamma^\tau_{\mu\nu} \frac{dx_\mu}{ds} \frac{dx_\nu}{ds} = 0. \quad (2)$$

In GR, material bodies subject only to gravitational fields are supposed to move on the geodesics determined by equation (2).³ The problem of motion in GR is the question of whether the equations of motion of matter subject to gravitational fields (2) can be derived from the gravitational field equations (1).

Einstein himself, in his first publication on the topic, a paper co-written with Jakob Grommer and published in 1927, compares different classes of attempts to give such a derivation. In particular, Einstein and Grommer distinguish between two classes of attempts at deriving the geodesic motion of matter from the gravitational field equations, which I will term *the T approach* and *the vacuum approach*, respectively. The *T* approach starts from the realization that the field equations (1) imply the conservation condition, namely that the covariant divergence of the energy-momentum tensor $T_{\mu\nu}$ vanishes:

$$\nabla^\mu T_{\mu\nu} = 0 \quad . \quad (3)$$

³It is a big question which systems are actually included under ‘material bodies’ here. The minimal position is that only test particles are referred to: particles with negligible extension, spin, and self-gravity. However, many actual bodies can be approximated well by test particles in this sense; planets orbiting a star are an example, as we shall see below.

From this, together with certain conditions on the energy-momentum tensor $T_{\mu\nu}$, the T approach derives that material particles move on time-like geodesics. It is this kind of approach to the problem of motion that philosophers have engaged with almost exclusively up to now.⁴

Einstein and Grommer end up dismissing the T approach, and suggest an alternative path to deriving geodesic motion instead. It is a particular version of a *vacuum approach to the problem of motion*. Einstein and Grommer start from the vacuum form of the Einstein field equations,

$$R_{\mu\nu} = 0 \quad , \quad (4)$$

and attempt to derive that the equations (4) imply that material particles move on geodesics.

To the extent that philosophers have engaged with this approach at all, they have quickly dismissed it because it seems to model material bodies by singularities in spacetime; while singularities, by definition, are not even part of spacetime. However, in this paper I shall argue that this dismissal was far too fast, and that indeed the vacuum approach deserves at least as much attention by philosophers as the T approach. The vacuum approach, despite first appearances, engages more closely with some of the most major predictions of GR: both the prediction of the perihelion of Mercury and the prediction of light bending by the Sun utilise the vacuum approach to the derivation of motion of material systems. Indeed, even the prediction of gravitational waves resulting from a binary black hole merger that was recently confirmed rests on the vacuum field equations, for black holes are described by vacuum solutions.⁵

My argument in this paper will proceed in three steps. First, I will argue that the vacuum approach to the problem of motion promises certain advantages that the T approach lacks. Second, I will argue that the problems of the vacuum approach for which it has been dismissed are artefacts of a too literal interpretation of the formalism and its application to the problem at hand. Third, I will argue that a careful interpretation makes the problems disappear; I will argue that the approach does not need to interpret singularities as representing material bodies.

⁴For a comprehensive review of the early history of this approach see Havas [1989] and Kennefick [2005]; for two particularly beautiful exemplars from within this class of proofs see Geroch and Jang [1975] and Ehlers and Geroch [2004], which are investigated by Brown [2007], Malament [2012], and Weatherall [Forthcoming, 2011].

⁵See Abbott et al. [2016] and references therein.

2 A critical comparison of the two research programmes

I said above that the T approach to the problem of motion proceeds via the fact that the Einstein field equations (1) imply the conservation condition (3), which in turn implies the geodesic motion of matter. However, as Malament [2012] pointed out, the conservation condition by itself is not sufficient to prove that the geodesic equation is the equation of motion of material particles. One of the most general proofs from within the T approach, proposed by Geroch and Jang [1975] and further generalised by Ehlers and Geroch [2004], rests not only on the conservation condition (3), but also on the strengthened dominant energy condition, which states:

Given any timelike covector ξ_μ at any point in M , $T^{\mu\nu}\xi_\mu\xi_\nu \geq 0$ and either $T^{\mu\nu} = \mathbf{0}$ or $T^{\mu\nu}\xi_\mu$ is timelike.

The first clause is effectively the weak energy condition, which states that the mass-energy-momentum density associated with the body in question is always non-negative. The second clause states that every observer will judge the mass-energy-momentum of the body to propagate along time-like curves only.⁶

It would be rather attractive if we did not have to presume that material particles move on time-like curves to then show that these curves are actually time-like geodesics, and if we did not have to presume that matter cannot have non-negative mass-energy. These are weak assumptions about the nature of matter, but they are assumptions.

The vacuum approach to the problem of motion, on the other hand, aims to make *no* assumptions about the nature of matter and its properties at all, and to still derive that matter moves on geodesics. It starts from the question of whether just knowing the exterior gravitational field of a material body, and how this gravitational field interacts with the gravitational field of its surroundings, is enough to derive that the body will move on a geodesic of the metric surrounding it. Arguably, this programme is far more ambitious than the *T* approach, for it starts with fewer assumptions.⁷ And yet, if successful, it would really fit much better the virtues that philosophers have associated

⁶For more on the interpretation of the strengthened dominant energy condition see Weatherall [2011], Weatherall [Forthcoming] and especially Curiel [Forthcoming].

⁷One might be tempted to argue that despite first appearances the vacuum approach

with the geodesic theorem(s) in the first place: deriving the inertial motion of matter from knowledge of the dynamics of gravitational fields alone.⁸

Einstein was deeply skeptical of the role of the energy-momentum tensor in GR. Throughout the decades, he emphasised that $T_{\mu\nu}$ provides only a ‘phenomenological representation of matter’.⁹ In Einstein and Grommer [1927], Einstein elaborates that general relativity with an energy-momentum tensor as a source term on the right-hand side of (1) is just not a complete theory: it does not tell us what kind of matter is present, only that it has a certain mass-energy distribution. This perspective on GR was further strengthened by Tupper [1981, 1982, 1983], who showed that knowing the energy-momentum tensor of a material system does not suffice to tell us what kind of matter is present. For example, one and the same mass-energy-momentum distribution $T_{\mu\nu}$ featuring on the right-hand side of the Einstein equations, and solving the Einstein equations for the same metric, can correspond either to an electromagnetic field or a viscous fluid. Knowing the energy-momentum tensor is just not sufficient to know which of these two material systems it is that interacts with the metric field.

Einstein’s aim is then to instead start with the vacuum field equations starts with more demanding assumptions than the T approach. For the vacuum Einstein equations (4) logically imply that the strengthened dominant energy condition (SDEC) holds for the Ricci tensor $R_{\mu\nu}$. The opposite is not true, so that demanding Ricci flatness is clearly a stronger constraint on the Ricci tensor than demanding that it obeys the SDEC. But concluding from this that the vacuum approach starts from stronger assumptions than the T approach would be a mistake. For the T approach assumes i.) the full Einstein field equations (1); and ii.) that the energy-momentum tensor (and thus the Einstein tensor) adheres to the SDEC. The vacuum approach only assumes the vacuum Einstein equations (4), and thus starts with weaker assumptions than the T approach. However, it might well be that despite *starting* with weaker assumptions than the T approach, a particular manifestation of the vacuum approach might end up with stronger assumptions than a particular manifestation of the T approach. For example, the 1927 Einstein-Grommer vacuum approach, discussed below, involves, among other demands, a so-called equilibrium condition which is supposed to relate solutions to the non-linear field equations to solutions of the linearized field equations in a particular way; no such demand is included in, say, the Geroch-Jang version of the T approach. Thus, further analysis might well show that Einstein and Grommer use stronger assumptions than Geroch and Jang. Einstein himself would likely have been content with that, as long as it allowed him to avoid the introduction of $T_{\mu\nu}$, for reasons discussed below.

⁸Cf. Brown [2007], p. 141 and 163.

⁹See, for example, Einstein [1922], Einstein to Michele Besso, 11 August 1926 (EA-7-361), and Einstein [1936].

(4), treat material particles as singularities in the metric field,¹⁰ and derive that they move on geodesics of a metric $g_{\mu\nu}$ that solves the vacuum field equations (4) in the region through which the particle moves.

To the extent that philosophers have engaged with this approach at all, they have already dismissed it at this point. The main criticism is that the very idea of the approach is flawed: A singularity is not even part of spacetime. How should it be possible to describe its motion in said spacetime?

Both Torretti and Earman essentially answer that this is not possible and that the whole programme is ill-conceived. Earman [1995], p. 12, writes:¹¹

[S]ingularities in the spacetime metric cannot be regarded as taking place at points of the spacetime manifold M . Thus, to speak of singularities in $g_{\mu\nu}$ as geodesics of the spacetime is to speak in oxymorons.

The most detailed discussion of the Einstein-Grommer paper in the philosophical literature is due to Tamir [2012]. After quoting the above statement by Earman, Tamir goes on to write (p.142):

The proponent of such a “vacuum-cum-singularity” technique is faced with the rather paradoxical challenge of explaining in what sense we can say that a singular curve (ostensibly constituted by the *missing* points in the manifold) is actually a geodesic of the spacetime from which it is absent. Not only is no metric defined at the singularity, but also technically there are not even spacetime points there: the geodesic does not exist.

Tamir then mentions a key ingredient of the Einstein-Grommer approach, namely the distinction between an ‘inner metric’ and an ‘outer metric’.¹² Einstein and Grommer aim to show that the particle characterized by a

¹⁰In recent years, the adequate definition of a singularity in GR has been a subject of extensive debate, see e.g. Earman [1995] and Curiel [1999]. For Einstein’s thoughts on singularities see Earman and Eisenstaedt [1999]; in the context of the Einstein-Grommer paper Einstein clearly thinks of a singularity in the metric field $g_{\mu\nu}$ as a region where the components of the metric tend to infinity.

¹¹For similar statements see Torretti [1996], section 5.8.

¹²There is an interesting relationship between Einstein and Grommer’s distinction between inner and outer metric (discussed further in section 3) on the one hand and the later distinction between interior and exterior black hole solutions on the other. I do believe that bringing together results and concepts developed in the context of black hole solu-

singular inner metric moves on geodesics of the non-singular outer metric. Tamir states that the “suggested implication” is that we are to compare a second spacetime whose metric is that of the regular outer metric with the singular first spacetime, and identify the regular geodesic of the second spacetime with the singular curve of the first one. He then argues that the thought that the second singularity-free spacetime can teach us anything about the singular original spacetime is “spurious”.

My point in the following will be this. Even if this argument were convincing, its premise (the ‘suggested implication’ that Einstein and Grommer intended to deduce something about a singular spacetime by comparing it to a non-singular spacetime) is not. I shall argue that by looking at the details of the Einstein-Grommer approach we come to a different interpretation of the approach, one that sheds a completely different light on the alleged presence of singularities. We will see that a careful (rather than literal) interpretation of the vacuum approach, and the Einstein-Grommer paper in particular, does not actually depend on introducing singularities at all.

3 The vacuum approach to the problem of motion

3.1 Two ways of looking at Einstein’s model of the Sun-Mercury system

In a way, the story of the vacuum approach to the problem of motion starts in 1915, with Einstein’s treatment of the orbit of Mercury around the Sun in the context of GR. It is a two-body problem: a small body (Mercury) with a comparatively small mass orbits a large body (the Sun). Einstein seems to postulate (more on the ‘seems’ below) that the Sun be represented by what would soon be recognized as an approximation to the Schwarzschild metric. He definitely postulates (!) that Mercury moves on a geodesic of said metric.¹³ In a way, the problem of motion in GR is about the question of

tions (a special case of vacuum solutions) on the one hand and the vacuum approach to the problem of motion on the other hand is very promising indeed. I will have to postpone a detailed discussion to a later paper; it will include the problem of motion of a binary black hole, the black hole equivalent of the Sun-Mercury two-body system discussed below.

¹³For a careful analysis of Einstein’s Mercury paper and how it rests on the Einstein-Besso manuscript see Earman and Janssen [1993], and Janssen’s Editorial Note on the

whether this second postulate is really necessary.

If we now look at Einstein's Mercury paper and recall the kind of criticism that was launched against the vacuum approach to the problem of motion, we may find ourselves feeling puzzled. After all, the Schwarzschild metric is a solution to the vacuum field equations, and it has a singularity at its center.¹⁴ If representing material bodies by singular metrics is so problematic, how does it come about that Einstein [1915] successfully predicted the perihelion motion of Mercury? Why is it not problematic to represent the Sun by the singular Schwarzschild metric?

The answer lies in denying the premise of the question. Einstein's treatment of the Sun-Mercury system should *not* be interpreted as involving him representing the Sun by (an approximation of) the Schwarzschild metric. We *know* that the Sun is a material body with non-vanishing mass-energy, and that it does not have a spacetime singularity at its center. What Einstein really does is to convert the two-body problem Sun-Mercury into a one-body problem, where one body (Mercury) is subject to an external gravitational field. It is the exterior gravitational field of the Sun, *not the Sun itself*, that is represented by the Schwarzschild metric. And that is enough to predict the perihelion of Mercury: we don't need to know what the Sun is made of or what happens in its interior; all that matters is the exterior gravitational field that Mercury is subject to.

Thus, worrying about the singularity at the center of the Schwarzschild metric just misses the point: we do not have to interpret the interior part of the Schwarzschild metric literally, at least not in this application.

In the following I shall argue that we should interpret the appearance of singularities in the Einstein-Grommer vacuum approach to the problem of motion in a similar vein.

3.2 The Einstein-Grommer vacuum approach to the problem of motion

The general scheme of the Einstein-Grommer approach proceeds as follows.¹⁵

Einstein-Besso manuscript in Vol. 4 of the Collected Papers of Albert Einstein (CPAE).

¹⁴For the history and interpretation of the Schwarzschild metric and its analytic extensions see Eisenstaedt [1989] and Bonnor [1992].

¹⁵The genesis of the Einstein-Grommer approach has been a bit of a mystery up to now, as pointed out by Kennefick [2005]. However, the work on the 15th volume of Einstein's collected papers has revealed the context and correspondence leading up to that paper,

1. Reformulate the vacuum Einstein equations in terms of a surface integral over a three-dimensional hyper-surface such that we can ask whether gravitational energy-momentum represented by the pseudo-tensor t_α^τ passes through the surface.¹⁶
2. Pick a curve that is supposed to represent the path of a material particle.
3. Impose the linear approximation according to which $g_{\mu\nu} = \eta_{\mu\nu} + \gamma_{\mu\nu}$, i.e. assume that, at least close to the curve, the metric deviates from Minkowski spacetime only slightly.
4. Realise that not all solutions to the linearized field equations will correspond to solutions of the non-linear field equations that the linearized field equations approximate. Argue that in the case where an ‘equilibrium condition’ for the energy-pseudo-tensor of the gravitational field holds, the $\gamma_{\mu\nu}$ of the linearized field equations *will* solve the full non-linear equations reformulated as a surface integral.¹⁷
5. Now split the $\gamma_{\mu\nu}$ in the immediate neighborhood of the particle into the ‘inner metric $\bar{\gamma}_{\mu\nu}$ that the particle itself gives rise to and the ‘outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ that is due to other sources (or lack thereof). Observe that the ‘outer metric’ is entirely regular, even if extended to the point at which the material particle is supposed to be located.
6. Integrate the surface integral that is equivalent to the vacuum field equations ‘around’ the curve that is supposed to represent the path of a material particle. For the case where the integration surface is a sphere, the equilibrium condition for t_α^τ simplifies to $\frac{\partial \bar{\bar{\gamma}}_{44}}{\partial x_\sigma} = 0$.

and how it fits into Einstein’s overall research program. It is a fascinating story; alas, it will have to wait for a separate paper.

¹⁶There has been a long debate on whether gravitational energy can be adequately represented by a pseudo-tensor; I will not be able to do it justice here. For some details see the introduction to Volume 8 CPAE for the debate between Einstein, Klein, Levi-Civita and Lorentz, for conceptual analysis Hofer [2000] and especially Trautmann [1962].

¹⁷This step is very intricate and it would take me a few pages to do it justice. This point of the Einstein-Grommer paper has not been addressed by the literature at all (neither in physics nor in philosophy); I will argue elsewhere that it sheds new light on Einstein’s later doubts as to whether the gravitational wave solutions of the linearized equations correspond to gravitational wave solutions in the full non-linear theory.

7. Conclude that the curve that represents the path of a material particle is a geodesic of the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$.¹⁸

4 Interpreting the Einstein-Grommer approach to the problem of motion

The reader might think that the argument presented in the last section cannot be a faithful representation of the Einstein-Grommer approach; after all, where is the claim that the material particle is represented by a singularity, the reason the approach was dismissed by Earman and Tamir? Indeed, I have omitted that after step 5 of the argument Einstein and Grommer *do* say that one *could* assume that the inner metric $\bar{\gamma}_{\mu\nu}$ is given by what is effectively a three-dimensional counterpart of the Schwarzschild metric: it is spherically symmetric and has a singularity at the center. And yet, *Einstein and Grommer never use this assumption in their argument*. They call the material particle ‘the singularity’ all the time, but their argument does not depend on assuming *any* particular form for the inner metric, let alone one that is necessarily singular. As a matter of fact, they do not even mention a concrete candidate metric for the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$; all they need is that $\gamma_{\mu\nu}$ is split into the inner metric $\bar{\gamma}_{\mu\nu}$ and the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ in such a way that $\bar{\bar{\gamma}}_{\mu\nu}$ is non-singular everywhere.

Note that this does not mean that we *know* that the inner metric $\bar{\gamma}_{\mu\nu}$ is non-singular. We don’t know anything about the inner metric, for the argument is independent of $\bar{\gamma}_{\mu\nu}$ having any particular form, just like the derivation of Mercury’s perihelion was independent of whether there is a singularity at the center of the Schwarzschild metric that represented the exterior field of the Sun.

With regard to the Sun-Mercury system I argued that we should not interpret the Schwarzschild metric as representing the Sun, but as representing its exterior gravitational field. The part of the Sun that is within the event horizon, including the singularity at the center, should not be taken

¹⁸Einstein and Grommer then go on to generalise this result to the ‘non-stationary case’, i.e. the case where it is not demanded that the external gravitational field, to which the particle is subject to, does not change in time. They conclude that in this case, too, the particle will move on a geodesic of the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ that is a solution to the field equations. For the following this generalisation does not make a difference; I will thus refer only to the stationary scenario described above.

as a representation of the *actual* interior of the Sun, but as a *placeholder* or a *blind spot* within the current description of the Sun-Mercury system: a docking station for a theoretical model of the Sun not included in Einstein's Sun-Mercury model.¹⁹

Likewise, we should interpret the inner metric $\bar{\gamma}_{\mu\nu}$ in the Einstein-Grommer approach as a placeholder for a representation of matter not included in the current theoretical approach. Sure, you *can* set $\bar{\gamma}_{\mu\nu}$ to be a Schwarzschild-like metric with a singularity at the center. But you don't have to do that to make the Einstein-Grommer argument work, and even if you do make that assumption, you should still take this particular inner metric with a singularity at its center as a placeholder for a representation or theory of matter not yet provided.²⁰

But now wait a minute. You might have disliked the occurrence of singularities as representations of particles, but at least the singularity (in lieu of a non-vanishing energy-momentum tensor) gave you an idea of *where* in spacetime the particle was supposed to be. True, Earman and Tamir rightly pointed out that the singularity is not actually part of spacetime, and so it can hardly serve to localize the particle in spacetime. Still, you might think that we're throwing the baby out with the bath water by not choosing any inner metric. After all, is it not the case then that the curve we have been focusing on is just *any* curve, without any reason to think of this curve as the curve of a material particle?²¹

Again, I think we can counter this criticism by comparing the Einstein-Grommer approach to Einstein's treatment of the Sun-Mercury system in

¹⁹Note that there are interior extensions of the Schwarzschild metric that model the interior of the Sun by solutions of the non-vacuum field equations (1), for example by an incompressible perfect fluid. See Bonnor [1992], section 5.

²⁰If I had given more historical details, I could have, I believe, shown that Einstein himself saw the occurrence of a singularity in the inner metric in exactly this way. This exegetical argument would have started with evidence that, from early on, he saw GR as a theory of the pure gravitational field without any constraints on what kinds of matter give rise to the gravitational field. Furthermore, I would have argued that even in the Einstein-Grommer paper he clearly forbids singularities *outside* of material particles (where the theory is supposed to give an adequate and deterministic representation of gravitational fields) but has no problem with them appearing *inside of* material systems, where the theory can provide at best phenomenological placeholders for a future 'proper' theory of matter anyhow. Thus, for Einstein energy-momentum tensors as alleged representatives of material systems were on a par with singularities: both were only placeholders for a proper theory of matter.

²¹I thank Jim Weatherall for putting this question to me.

Einstein [1915]. What Einstein did there was to assume that Mercury would move on *some* geodesic of the exterior gravitational field produced by the Sun. He calculated an approximation to the external gravitational field of a static, spherically symmetric and asymptotically flat body; this gravitational field he saw as represented by the connection components $\Gamma^\nu_{\mu\sigma}$ of a metric $g_{\mu\nu}$ which deviated only slightly from the flat Minkowski metric. He then inserted these gravitational field components $\Gamma^\nu_{\mu\sigma}$ into the geodesic equation (2). He showed that this law contained Newton's first law and Newton's second law with a gravitational potential giving rise to a force as a limiting case, and showed how the resulting Keplerian laws for orbits differ in his theory as compared to its Newtonian limit. In the end, he obtained that according to the new theory the perihelion ϵ of *any* geodesic orbit around the Sun is given by

$$\epsilon = 24\pi^3 \frac{a^2}{T^2 c^2 (1 - e^2)} \quad (5)$$

Here a denotes the length of the semimajor axis of the orbit in question, e its eccentricity, c the speed of light, and T the orbital period of the planet in question. Einstein then *takes the astronomically known values for Mercury*, plugs them into equation (5), and thereby predicts that Mercury's perihelion changes by 43" per century.

Note that there is *nothing* in the theoretical description that singles out any particular path as that of Mercury. There is no theoretical representation of Mercury, no model. All that is there is the assumption that Mercury will move on one of the geodesics of the affine connection determined by the spherically symmetric field of the Sun. A general equation that all possible geodesic orbits have to fulfil is derived. And then *external knowledge* is used to single out one of these orbits as that of Mercury. Einstein trusts that the astronomers have measured the orbital period, the semimajor axis and the eccentricity of Mercury correctly. It is this external knowledge, plugged into his theoretical model, which does not in itself contain a representation of Mercury or its path, that produces the prediction.

In many ways, the whole vacuum approach to the problem of motion is about the question as to whether in this kind of scenario we really have to assume the geodesic equation as the equation of motion of matter over and above the gravitational field equations. Indeed, let us look at the Sun-Mercury system within the 1927 Einstein-Grommer approach. The problem of motion, then, is the question whether Einstein really *had to* introduce the

gravitational field equations (to describe the exterior gravitational field of the Sun) *and* the geodesic equation (to describe the path of Mercury subject to this gravitational field) as separate assumptions.²² Could he have only assumed the gravitational field equations and *derived* that Mercury moves on a geodesic of the exterior field of the Sun? My point is that, just like in Einstein's 1915 treatment, the 1927 Einstein-Grommer approach does not *need to* commit to a theoretical model that allows us to localise Mercury internally. It is fine to ask whether the exterior gravitational field around a given curve 'forces' that curve to be a geodesic. Just like in the 1915 treatment, Einstein and Grommer could then use *external knowledge* about whether that particular curve is actually the curve of a material object, or of Mercury in particular. No inner metric, no singularity to represent the material body, is actually needed.

Let us take a step back though, for there is an important difference between the structure of Einstein's 1915 treatment of Mercury on the one hand and the 1927 Einstein-Grommer approach on the other. In the Mercury case Einstein had assumed (!) that Mercury moves on a geodesic, i.e. a special kind of curve, and model-external knowledge about the period, eccentricity and semimajor axis of Mercury could then be used to determine which of the many geodesics of the Schwarzschild metric corresponded to the path of Mercury. But in the case of the Einstein-Grommer argument, what is in question is whether we can prove that the path of Mercury, say, is a geodesic. Thus, at first sight it looks as if while the 1915 argument only needed external knowledge to determine which geodesic is that of Mercury, appeal to external knowledge in the Einstein-Grommer case would have to determine a.) that this curve is a geodesic and b.) that it is the curve of a material body.

Einstein and Grommer did not aim to derive both a.) and b.). Instead, while Einstein in 1915 used external knowledge at the end of his argument, Einstein and Grommer in 1927 use it at the beginning. They start out by assuming that a given curve is the curve of a material particle, and then ask whether having a regular outer metric (which solves the vacuum field equations) around the curve means that the curve of this material particle,

²²Interestingly, Einstein did not yet have the final gravitational field equations in the Mercury paper; he found them a week later, in his fourth paper of November 1915. However, the approximation of the Schwarzschild metric that he uses in the Mercury paper is an approximative solution of both the field equations from the Mercury paper, and of the final Einstein field equations.

given the further conditions summarized in section 3.2, *must be* a geodesic. Rather than finishing the argument by appeal to external knowledge (as in Einstein 1915), the Einstein-Grommer argument starts with an appeal to external knowledge, which singles out a particular curve as that of a material body.²³

Either way, both in Einstein's 1915 treatment and in the Einstein-Grommer approach there is no reason to interpret the singularity (appearing in the Schwarzschild metric or the inner metric, respectively) literally. In both cases, the singularity should be interpreted to signify a placeholder or a blind spot of the theoretical treatment, rather than something that should be interpreted literally, as referring and approximately true. Indeed, both Einstein's 1915 treatment of the Sun-Mercury system and Einstein's and Grommer's treatment of an arbitrary material particle subject to an external gravitational field work just as well if, in the former case, no interior metric (to describe the interior of the Sun) or, in the latter case, no inner metric (to represent the location of the particle on the curve), is ever specified.

5 Conclusion

I started out by saying that whether we are realists or antirealists, we should aim for a careful interpretation, rather than a literal interpretation, of the scientific theory that we want to be realists or anti-realists about. As a case study, I argued that the vacuum approach to the problem of motion in GR, and the Einstein-Grommer approach in particular, is far more sensible and promising if we interpret the singularities *not as representing* material bodies but as *placeholders* for a representation of material bodies that is not included in the model. Indeed, I argued that the approach does not even need the

²³There is a further disanalogy between Einstein's 1915 derivation of the perihelion of Mercury and the Einstein-Grommer argument of 1927. In the former the choice of (an approximation) the Schwarzschild metric to represent the exterior gravitational field of the Sun does important work in the derivation of Mercury's perihelion. In the Einstein-Grommer approach, no choice of a concrete outer metric is necessary to derive that the curve of the particle which is surrounded by the outer metric must be a geodesic. The reason for this difference is that the Einstein-Grommer approach aims to be more general; it only aims to derive *that* a material body moves on *some* geodesic of the outer metric. However, note that it is not the case that any outer metric is allowed by the approach: the class of outer metrics that the approach can work with is heavily constrained by steps 2 and 3 of the Einstein-Grommer argument (see section 3.2).

introduction of singularities to represent material bodies; their introduction does not do any work in answering the question at hand.²⁴

Given that in their paper Einstein and Grommer seem to take the singularities as representing material bodies, one might wonder whether this allegedly more careful interpretation does not fall prey to the criticism that the careful interpreter presumes to understand the theory/formalism in question better than its originators. This might seem at odds with the realist tenet of taking scientists and science ‘seriously’. I do indeed think that putting the Einstein-Grommer paper into its proper historical context by analysing Einstein’s correspondence leading up to the paper and by relating it to his overarching research project at the time *would* convincingly show that he subscribed to something very much like the ‘placeholder interpretation’ I defended above. Showing this in detail will have to wait for a much longer paper, and I do not ask the reader to just take my word for it. So let us say, for the sake of the argument, that Einstein and Grommer did indeed intend the singularities as representatives of material objects in a rather straightforward way. I believe that we should not take *their* word for it either. And neither did Einstein. Just a few years after the Einstein-Grommer paper, in his famed 1933 Spencer lectures at the University of Oxford, Einstein told us in his opening words: “If you wish to learn from the theoretical physicist anything about the methods which he uses, I would give you the following advice: Don’t listen to his words, examine his achievements.”²⁵

In philosophy of science, I believe there is no better way of examining a scientist’s achievements than by looking for the best possible interpretation

²⁴The argument that we should thus not see a realist as committed to being a realist *about* the singularities appearing in the Einstein-Grommer paper resonates well with selective or posit realism as introduced by Vickers [2013]. The idea there is that we should only be realists with respect to components of a prediction that ‘fuel the success’ of the prediction, i.e. that are indispensable in the derivation of what is predicted. Using Vickers’ distinction the introduction of a singular inner metric in the Einstein-Grommer approach is an idle rather than a working posit. However, note that the call for careful rather than literal interpretations with which I started is independent of / complementary to aiming for identification of the idle posits in a derivation. For *even if* we had found that the introduction of the singular inner metric did do work in the derivation of geodesic motion could we have argued (with less force) that the singularity should be interpreted as a placeholder for a future theory of matter, as a temporary measure within an effective theory, and thus not as something that we should interpret as possessing as much ‘reality’ or ‘referring power’ as the regular outer metric governed by the field equations.

²⁵See Einstein [1934], and van Dongen [2010] for a detailed analysis of the text.

of his or her theories. To do that, we have to not just listen to the words of the scientist who created or discovered it; we have to see what the theory *does* in practice, how it is *used*; which of its parts really do the work.

Acknowledgments

I would like to thank my colleagues at Caltech and at the Einstein Papers Project for many discussions about the problem of motion and the Einstein-Grommer approach in particular. Thanks are due especially to Diana Kormos-Buchwald, Frederick Eberhardt, and Daniel Kennefick. I would also like to thank audiences at Caltech, Oxford, Irvine, the BSPS 2016 conference in Cardiff, and at the 8th Quadrennial Pittsburgh Fellows conference in Lund, Sweden for many helpful discussions on the topic. I would like to thank especially Sam Fletcher, David Malament and Jim Weatherall for carefully reading earlier versions of this paper, and for the extremely helpful comments they gave me. Finally, I would like to thank Dana Tulodziecki for pointing my attention to the link between posit realism and what I was saying in this paper.

References

- Abbott, B., Abbott, R., Abbott, T., Abernathy, M., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R. et al. [2016], ‘Observation of gravitational waves from a binary black hole merger’, *Physical Review Letters* **116**(6), 061102.
- Bonnor, W. [1992], ‘Physical interpretation of vacuum solutions of einstein’s equations. part i. time-independent solutions’, *General relativity and Gravitation* **24**(5), 551–574.
- Brown, H. R. [2007], *Physical Relativity. Space-time structure from a dynamical perspective*, Oxford University Press, USA.
- Curiel, E. [1999], ‘The analysis of singular spacetimes’, *Philosophy of Science* pp. S119–S145.
- Curiel, E. [Forthcoming], A primer on energy conditions, *in* D. Lehmkuhl,

- G. Schiemann and E. Scholz, eds, ‘Towards a Theory of Spacetime Theories’, Einstein Studies, Birkhäuser.
- Earman, J. [1995], *Bangs, crunches, whimpers, and shrieks: Singularities and acausalities in relativistic spacetimes*, Oxford University Press, USA.
- Earman, J. and Eisenstaedt, J. [1999], ‘Einstein and singularities’, *Studies in History and Philosophy of Modern Physics* **30**(2), 185–235.
- Earman, J. and Janssen, M. [1993], ‘Einstein’s explanation of the motion of mercury’s perihelion’, *Einstein Studies* pp. 129–172.
- Ehlers, J. and Geroch, R. [2004], ‘Equation of motion of small bodies in relativity’, *Annals of Physics* **309**, 232–236.
- Einstein, A. [1915], ‘Erklärung der perihelbewegung des merkur aus der allgemeinen relativitätstheorie’, *Königliche Preussische Akademie der Wissenschaften (Berlin)* .
- Einstein, A. [1922], *Vier Vorlesungen über Relativitätstheorie gehalten im Mai 1921 an der Universität Princeton*, F. Vieweg. Reprinted as Vol.7, Doc. 71 CPAE; and in various editions as “The Meaning of Relativity” by Princeton University Press.
- Einstein, A. [1934], ‘On the method of theoretical physics’, *Philosophy of science* **1**(2), 163–169.
- Einstein, A. [1936], ‘Physics and reality’, *Journal of the Franklin Institute* **221**, 349–382. Reprinted in A. Einstein (1976) *Ideas and Opinions* (New York: Dell Publishers), pp. 283-315.
- Einstein, A. and Grommer, J. [1927], ‘Allgemeine Relativitätstheorie und Bewegungsgesetz’, *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse* pp. 2–13.
- Eisenstaedt, J. [1989], The early interpretation of the schwarzschild solution, in D. H. a. J. Stachel, ed., ‘Einstein and the History of General Relativity’, Birkhäuser, pp. 1–213.
- Geroch, R. and Jang, P. [1975], ‘Motion of a body in general relativity’, *Journal of Mathematical Physics* **16**, 65–67.

- Havas, P. [1989], The early history of the "problem of motion" in general relativity, in 'Einstein and the History of General Relativity', Vol. 1, pp. 234–276.
- Hofer, C. [2000], 'Energy conservation in gtr', *Studies in History and Philosophy of Modern Physics* **31**.
- Kennefick, D. [2005], 'Einstein and the problem of motion: a small clue', *The universe of general relativity* pp. 109–124.
- Lehmkuhl, D. [2014], 'Why einstein did not believe that general relativity geometrizes gravity', *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **46**, 316–326.
- Malament, D. B. [2012], A remark about the "geodesic principle" in general relativity, in 'Analysis and Interpretation in the Exact Sciences', Springer, pp. 245–252.
- Pickering, A. [1999], *Constructing quarks: A sociological history of particle physics*, University of Chicago Press.
- Saunders, S., Barrett, J., Kent, A. and Wallace, D. [2010], *Many worlds? Everett, quantum theory, & reality*, OUP Oxford.
- Tamir, M. [2012], 'Proving the principle: Taking geodesic dynamics too seriously in Einstein's theory', *Studies In History and Philosophy of Modern Physics* **43**(2), 137–154.
- Torretti, R. [1996], *Relativity and geometry*, Dover Publications.
- Trautmann, A. [1962], Conservation laws in general relativity, in L. Witten, ed., 'Gravitation: An Introduction to Current Research', John Wiley and Sons.
- Tupper, B. [1981], 'The equivalence of electromagnetic fields and viscous fluids in general relativity', *Journal of Mathematical Physics* **22**(11), 2666–2673.
- Tupper, B. [1982], 'The equivalence of perfect fluid space-times and magnetohydrodynamic space-times in general relativity', *General Relativity and Gravitation* **15**(1).

- Tupper, B. [1983], ‘The equivalence of perfect fluid space-times and viscous magnetohydrodynamic space-times in general relativity’, *General Relativity and Gravitation* **15**(9).
- van Dongen, J. [2010], *Einstein’s Unification*, Cambridge University Press, Cambridge.
- Van Fraassen, B. C. [1980], *The scientific image*, Oxford University Press.
- Vickers, P. [2013], ‘A confrontation of convergent realism’, *Philosophy of Science* **80**(2), 189–211.
- Weatherall, J. O. [2011], ‘On the status of the geodesic principle in newtonian and relativistic physics’, *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics* **42**(4), 276 – 281.
URL: <http://www.sciencedirect.com/science/article/pii/S1355219811000566>
- Weatherall, J. O. [Forthcoming], Inertial motion, explanation, and the foundations of classical spacetime theories, in D. Lehmkuhl, G. Schieman and E. Scholz, eds, ‘Towards a Theory of Spacetime Theories’, Birkhäuser.

**Holism, or the Erosion of Modularity –
a Methodological Challenge for Validation**

Draft to be presented at PSA 2016

Johannes Lenhard, Bielefeld University

abstract

Modularity is a key concept in building and evaluating complex simulation models. My main claim is that in simulation modeling modularity degenerates for systematic methodological reasons. Consequently, it is hard, if not impossible, to access how representational (inner mathematical) structure and dynamical properties of a model are related. The resulting problem for validating models is one of holism.

The argument will proceed by analyzing the techniques of parameterization, tuning, and kludging. They are – to a certain extent – inevitable when building complex simulation models, but corrode modularity. As a result, the common account of validating simulations faces a major problem and testing the dynamical behavior of simulation models becomes all the more important. Finally, I will ask in what circumstances this might be sufficient for model validation.

1. Introduction

For the moment, imagine a scene at a car racing track. The air smells after gasoline. The pilot of the F1 racing car has just steered into his box and is peeling himself out of the straight cockpit. He puts off his helmet, shakes his sweaty hair, and then his eyes make contact to the technical director with a mixture of anger, despair, and helplessness. The engine had not worked as it should, and for a known reason: the software. However, the team had not been successful in attributing the miserable performance to a particular parameter setting. The machine and the software interacted in unforeseen and intricate ways. This explains the exchange of glances between pilot and technical director. The software's internal interactions and interfaces proved to be so complicated that the team had not been able to localize an error or a bug, rather remained

suspicious that some complex interaction of seemingly innocent assumptions or parameter settings was leading to the insufficient performance.

The story happened in fact¹ and it is remarkable since it displays how invasive computational modeling is into areas that smell most analogous. I reported this short piece for another reason, however, namely because the situation is typical for complex computational and simulation models. Validation procedures, while counting on modularity, run against a problem of holism.

Both concepts, modularity and holism, are notions at the fringe of philosophical terminology. Modularity is used in many guises and is not a particularly philosophical notion. It features prominently in the context of complex design, planning, and building – from architecture to software. Modularity stands for first breaking down complicated tasks into small and well-defined sub-tasks and then re-assembling the original global task with a well-defined series of steps. It can be argued that modularity is the key pillar on which various rational treatments of complexity rest – from architecture to software engineering.

Holism is a philosophical term to a somewhat higher degree and is covered in recent compendia. The Stanford Encyclopedia, for instance, includes (sub-)entries on methodological, metaphysical, relational, or meaning holism. Holism generically states that the whole is greater than the sum of its parts, meaning that the parts of a whole are in intimate interconnection, such that they cannot exist independently of the whole, or cannot be understood without reference to the whole. Especially W. V. O. Quine has made the concept popular, not only in philosophy of language, but also in philosophy of science, where one speaks of the so-called Duhem-Quine thesis. This thesis is based on the insight that one cannot test a single hypothesis in isolation, but that any such test depends on “auxiliary” theories or hypotheses, for example how the measurement instruments work. Thus any test addresses a whole ensemble of theories and hypotheses.

Lenhard and Winsberg (2010) have discussed the problem of confirmation holism in the context of validating complex climate models. They argued that “due to interactivity, modularity does not break down a complex system into separately manageable pieces.” (2010, 256) In a sense, I want to pick up on this work, but put the thesis into a much more general context, i.e. pointing

¹ In spring 2014, the Red Bull team experienced a crisis due to recalcitrant problems with the Renault engine, due to a partial software update.

out a dilemma that is built on the tension between modularity and holism and that occurs quite generally in simulation modeling. The potential philosophical novelty is debated controversially in philosophy of science, for instance Humphreys (2009) vs. Frigg and Reiss (2009). The latter authors deny novelty, but concede issues of holism might be an exception. My paper confirms that holism is a key concept when reasoning about simulation. (I see more reasons for philosophical novelty, though.)

My main claim is the following: According to the rational picture of design, modularity is a key concept in building and evaluating complex models. In simulation modeling, however, modularity erodes for systematic methodological reasons. Moreover, the very condition for success of simulation undermines the most basic pillar of rational design. Thus the resulting problem for validating models is one of (confirmation) holism.

Section 2 discusses modularity and its central role for the so-called rational picture of design. Herbert Simon's highly influential parable of the watchmakers will feature prominently. It paradigmatically captures complex systems as a sort of large clockwork mechanism. This perspective suggests the computer would enlarge the tractability of complex systems due to its vast capacity for handling (algorithmic) mechanisms. Complex simulations then would appear as the electronic incarnation of a gigantic assembly of cogwheels. This viewpoint is misleading, I will argue. Instead, I want to emphasize the dis-analogy to how simulation models work. The methodology of building complex simulation models thwarts modularity in systematic ways. Simulation is based on an iterative and exploratory mode of modeling that leads to a sort of *holism that erodes modularity*.

I will present two arguments for the erosion claim, one from parameterization and tuning (section 3), the other from klu(d)ging (section 4). Both are, in practice, part-and-parcel of simulation modeling and both make modularity erode. The paper will conclude by drawing lessons about the limits of validation (section 5). Most accounts of validation require, if often not explicitly, modularity and are incompatible with holism. In contrast, the exploratory and iterative mode of modeling restricts validation, at least to a certain extent, to testing (global) predictive virtues. This observation shakes the rational (clockwork) picture of design and of the computer.

2. The rational picture

The design of complex systems has a long tradition in architecture and engineering. At the same time, it has not been much covered in literature, because design was conceived as a matter for experienced craftsmanship rather than analytical investigation. The work of Pahl and Beitz (1984, plus revised editions 1996, 2007) gives a relatively recent account of design in engineering. A second, related source for reasoning about design is the design of complex computer systems. Here, one can find more explicit accounts, since the computer led to complex systems much faster than any tradition of craftsmanship could grow. A widely read example is Herbert Simon's "Sciences of the Artificial" (1969). Still up to today, techniques of high-level languages, object-oriented programming, etc. make the practice of design change on a fast scale.

One original contributor to this discussion is Frederic Brooks, software and computer expert (and former manager at IBM) and also hobby architect. In his 2010 monograph "The Design of Design", he describes the rational model of design that is widely significant, though it is much more often adopted in practice than explicitly formulated in theoretical literature. The rational picture starts with assuming an overview of all options at hand. According to Simon, for instance, the theory of design is the general theory of search through large combinatorial spaces (Simon 1969, 54). The rational model then presupposes a utility function and a design tree, which are spanning the space of possible designs. Brooks rightly points out that these are normally not at hand. Nevertheless, design is conceived as a systematic step-by-step process. Pahl and Beitz aim at detailing these steps in their rational order. Also, Simon presupposes the rational model, arguably motivated by making design feasible for artificial intelligence (see Brooks 2010, 16). Wynston Royce, to give another example, introduced the "waterfall model" for software design (1970). Royce was writing about managing the development of large software systems and the waterfall model consisted in following a hierarchy ("downward"), admitting to iterate steps on one layer, but not with much earlier ("upward") phases of the design process. Although Royce actually saw the waterfall model as a straw man, it was cited positively as paradigm of software development (cf. Brooks on this point).

Some hierarchical order is a key element of the rational picture of design and presumes modularity. Let me illustrate this point. Consider first a simple brick wall. It consists of a multitude of modules, each with certain form and static properties. These are combined into

potentially very large structures. It is a strikingly simple example, because all modules (bricks) are similar.

A more complicated, though closely related, example is the one depicted in figure 1 where an auxiliary building of Bielefeld University is put together from container modules.



Figure 1: A part of Bielefeld University is built from container modules.

These examples illustrate how deeply ingrained modularity is in our way of building (larger) objects. Figure 2 displays a standard picture for designing and developing complex (software) systems.

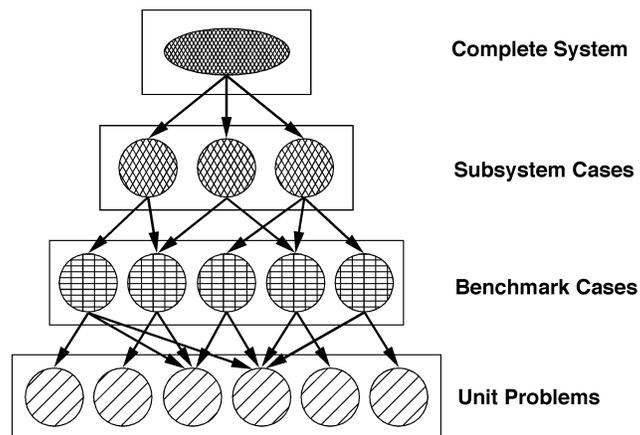


Figure 2: Generic architecture of complex software, from the AIAA Guide for the Verification and Validation of Computational Fluid Dynamics Simulations (1998). Modules of one layer might be used by different modules on a higher layer.

Some complex overall task is split up into modules that can be tackled independently and by different teams. The hierarchical structure shall ensure the modules can be integrated to make up the original complex system. Modularity not only plays a key role when designing and building complex systems, it also is of crucial importance when taking account of the system. Validation is usually conceived in the very same modular structure: independently validated modules are put together in a controlled way for making up a validated bigger system. The standard account of how computational models are verified and validated gives very rigorous guidelines that are all based on the systematic realization of modularity (Oberkampf and Roy 2010, see also Fillion 2017). In short, modularity is key for designing as well as for validating complex systems.

This observation is paradigmatically expressed in Simon's parable of the two watchmakers. You find it in Simon's 1962 paper "The Architecture of Complexity" that has become a chapter in his immensely influential "The Sciences of the Artificial" (Simon 1969). There, Simon investigates the structure of complex systems. The stable structures, so Simon argues, are the hierarchical ones. He expressed his idea by telling the parable of the two watchmakers named Hora and Tempus (1969, 90-92). P. Agre describes the setting with the following words:

"According to this story, both watchmakers were equally skilled, but only one of them, Hora, prospered. The difference between them lay in the design of their watches. Each design involved 1000 elementary components, but the similarity ended there. Tempus' watches were not hierarchical; they were assembled one component at a time. Hora's watches, by contrast, were organized into hierarchical subassemblies whose "span" was ten. He would combine ten elementary components into small subassemblies, and then he would combine ten subassemblies into larger subassemblies, and these in turn could be combined to make a complete watch." (Agre 2003)

Since Hora takes additional steps for building modules, Tempus' watches need less time for assembly. However, it was Tempus' business that did not thrive, because of an additional condition not yet mentioned, namely some kind of noise. From time to time the telephone rings and whenever one of the watchmakers answers the call, all cogwheels and little screws fall apart and he has to re-start the assembly. While Tempus had to start from scratch, Hora could keep all finished modules and work from there. In the presence of noise, so the lesson goes, the modular

strategy is by far superior. Agre summarizes that modularity, he speaks of the functional role of components, comes out as a necessary element when designing complex systems:

“For working engineers, hierarchy is not mainly a guarantee that subassemblies will remain intact when the phone rings. Rather, hierarchy simplifies the process of design cognitively by allowing the functional role of subassemblies to be articulated in a meaningful way in terms of their contribution to the function of the whole. Hierarchy allows subassemblies to be modified somewhat independently of one another, and it enables them to be assembled into new and potentially unexpected configurations when the need arises. A system whose overall functioning cannot be predicted from the functionality of its components is not generally considered to be well-engineered.” (Agre 2003)

Now, the story works with rather particular examples insofar as watches exemplify complicated mechanical devices. The universe as a giant clockwork has been a common metaphor since the seventeenth century. Presumably, Simon was aware the clockwork picture is limited and he even mentioned that complicated interactions could lead to a sort of pragmatic holism.² Nonetheless, the hierarchical order is established by the interaction of self-contained modules.

There is an obvious limit to the watchmaker picture, namely systems have to remain manageable by human beings (watchmakers). There are many systems of practical interest that are too complex – from the earth’s climate to the aerodynamics of an airfoil. Computer models open up a new path here, since simulation models might contain a wealth of algorithmic steps far beyond what can be conceived in a clockwork picture.³ From this point of view, the computer appears as a kind of amplifier that helps to revitalize the rational picture. Do we have to look at simulation models as a sort of gigantic clockworks? In the following, I will argue that this viewpoint is seriously misleading. Simulation models are different from watches in important ways and I

² This kind of holism hence can occur even when modules are “independently validated”, since these modules when connected together could interact with each other in unpredicted ways. This is a strictly weaker form of holism than the one I am going to discuss.

³ Charles Babbage had designed his famous „Analytical Engine“ as a *mechanistic* computer. Tellingly, it did encounter serious problems exactly because of the mechanical limitations of its construction.

want to focus on the dis-analogy.⁴ Finally, we will learn from the investigation of simulation models about our picture of rationality.

3. Erosion of modularity 1: Parameterization and tuning

In stark contrast to the cogwheel picture of the computer, the methodology of simulation modeling erodes modularity in systematic ways. I want to discuss two separate though related aspects, firstly, parameterization and tuning and, secondly, kludging (also called kludging). Both are, for different reasons, part-and-parcel of simulation modeling; and both make modularity of models erode. Let us investigate them in turn and develop two arguments for erosion.

Parameterization and tuning are key elements of simulation modeling that stretch the realm of tractable subject matter much beyond what is covered by theory. Furthermore, simulation models can make predictions even in fields that *are* covered by well-accepted theory only with the help of parameterization and tuning. In this sense, the latter are success conditions for simulations.

Before we start with discussing an example, let me add a few words about terminology. There are different expressions that specify what is done with parameters. The four most common ones are (in alphabetical order): adaptation, adjustment, calibration, and tuning. These notions describe very similar activities, but also value differently what parameters are good for. Calibration is commonly used in the context of preparing an instrument, like calibrating a scale one time for using it very often in a reliable way. Tuning has a more pejorative tone, like achieving a fit with artificial measures, or fitting to a particular case. Adaptation and adjustment have more neutral meanings.

Atmospheric circulation is a typical example. It is modeled on the basis of accepted theory (fluid dynamics, thermodynamics, motion) on a grand scale. Climate scientists call this the “dynamical core” of their models and there is more or less consensus about this part. Although the employed theory is part of physics, climate scientists mean a different part of their models when they speak of “the physics”. It includes all the processes that are not completely specified from the dynamical core. These processes include convection schemes, cloud dynamics, and many more.

⁴ There are several dis-analogies. One I am not discussing is that clockworks lack multi-functionality.

The “physics” is where different models differ and the physics is what modeling centers regard as their achievements and try to maintain even if their models change into the next generation.

The physics acts like a specifying supplement to the grand scale dynamics. It is based on modeling assumptions, say which sub-processes are important in convection, what should be resolved in the model and what should be treated via a parameterization scheme. Often, such processes are not known in full detail, and some aspects (at least) depend on what happens on a sub-grid scale. The dynamics of clouds, for instance, depends on a staggering span of very small (molecular) scales and much larger scales of many kilometers. Hence even if the laws that guide these processes would be known, they could not be treated explicitly in the simulation model. Modeling the physics has to bring in parameterization schemes.⁵

How does moisture transport, for example, work? Rather than trying to investigate into the molecular details of how water vapor is entrained into air, scientists use a parameter, or a scheme of parameters, that controls moisture uptake so that known observations are met. Often, such parameters do not have a direct physical interpretation, nor do they need one, like when a parameter stands for a mixture of processes not resolved in the model. The important property rather is that they make the parameterization scheme flexible, so that the parameters of such a scheme can be changed in a way that makes the properties of the scheme (in terms of climate dynamics) match some known data or reference points.

From this rather straightforward observation follows an important fact. A parameterization, including assignments of parameter values, makes sense only in the context of the larger model. Observational data are not compared to the parameterization in isolation. The Fourth Assessment Report of the IPCC acknowledges the point that “parameterizations have to be understood in the context of their host models” (Solomon et al. 2007, 8.2.1.3)

The question of whether the parameter value that controls moisture uptake (in our oversimplified example) is adequate can be answered only by examining how the entire parameterization behaves and, moreover, how the parameter value in the parameterization in the larger simulation model behaves. Answering such questions would require, for instance, looking at more global properties like mean cloud cover in tropical regions, or the amount of rain in some area. Briefly

⁵ Parameterization schemes and their more or less autonomous status are discussed in the literature, cf. Parker 2013, Smith 2002, or Gramelsberger and Feichter 2011.

stated, parameterization is a key component of climate modeling and tuning is part-and-parcel of parameterization.⁶

It is important to note that tuning one parameter takes the values of other parameters as given, be they parameters from the same scheme, or be they parts of other schemes that are part of the model. A particular parameter value (controlling moisture uptake) is judged according to the results it yields for the overall behavior (like cloud cover). In other words, tuning is a local activity that is oriented at global behavior. Researchers might try to optimize parameter values simultaneously, but for reasons of computational complexity, this is possible only with a rather small subset of all parameters. A related issue is statistical regression methods that might be caught up in a local optimum. In climate modeling, skill and experience remain to be important for tuning (or adjustment).

Furthermore, tuning parameters is not only oriented at the global model performance, it tends to blur the local behavior. This is because every model will be importantly imperfect, since it contains technical errors, works with insufficient knowledge, etc. – which is just the normal case in scientific practice. Now, tuning a parameter according to the overall behavior of the model then means that the errors, gaps, and bugs get compensated against each other (if in an opaque way). Mauritsen et al. (2012) have pointed this out in their pioneering paper about tuning in climate modeling.

In climate models, cloud parameterizations play an important role, because they influence key statistics of the climate and, at the same time, cover major (remaining) uncertainties about how an adequate model should look like. Typically, such a parameterization scheme includes more than two dozens of parameters; most of them do not carry a clear physical interpretation. The simulation then is based on the balance of these parameters in the context of the overall model (including other parameterizations). Over the process of adjusting the parameters, these schemes become inevitably convoluted. I leave aside that models of atmosphere and oceans get coupled, which arguably aggravates the problem.

⁶ The studies of so-called perturbed physics ensembles convincingly showed that crucial properties of the simulation models hinge on exactly how parameter values are assigned (Stainforth et al. 2007).

Tuning is inevitable, part-and-parcel of simulation modeling methodology. It poses great challenges, like finding a good parameterization scheme for cloud dynamics, which is a recent area of intense research in meteorology. But when is a parameterization scheme a good one? On the one side, a scheme is sound when it is theoretically well motivated, on the other side, the key property of a parameterization scheme is its adaptability. Both criteria do not point into the same direction. There is, therefore, no optimum; finding a balance is still considered an art. I suspect that the widespread reluctance against publishing about practices of adjusting parameters comes from reservations against aspects that call for experience and art rather than theory and rigorous methodology.

I want to maintain that nothing in the above argumentation is particular to climate. Climate modeling is just one example out of many. The point holds for simulation modeling quite generally. Admittedly, climate might be a somewhat peculiar case, because it is placed in a political context where some discussions seem to require that only ingredients of proven physical justification and realistic interpretation are admitted. Arguably, this expectation might motivate using the pejorative term of tuning. This reservation, however, ignores the very methodology of simulation modeling. Adjusting parameters is by no means particular to climate modeling, nor is it confined to areas where knowledge is weak.

Another example will document this. Adjusting parameters is also occurring thermodynamics, an area of physics with very high theoretical reputation. The ideal gas equation is even taught in schools, it is a so-called equation of state (EoS) that describes how pressure and temperature depend on each other. However, actually using thermodynamics requires to work with less idealized equations of state than the ideal gas equation. More complicated equations of state find wide applications also in chemical engineering. They are typically very specific for certain substances and require extensive adjustment of parameters as Hasse and Lenhard (2017) describe and analyze. Clearly, being able to process specific adjustment strategies that are based on parameterization schemes is a crucial success condition. Simulation methods have made applicable thermodynamics in many areas of practical relevance, exactly because equations of state are tailored to particular cases of interest via adjusting parameters.

One further example is from quantum chemistry, namely the so-called density functional theory (DFT), a theory developed in the 1960s that won the Nobel prize in 1998. Density functionals

capture the information of the Schroedinger equation, but are much more computationally tractable. However, only many-parameter functionals brought success in chemistry. The more tractable functionals with few parameters worked only in simpler cases of crystallography, but were unable to yield predictions accurate enough to be of chemical interest. Arguably, being able to include and adjust more parameters has been the crucial condition that had to be satisfied before DFT could gain traction in computational quantum chemistry, which happened around 1990. This traction, however, is truly impressive. DFT is by now the most widely used theory in scientific practice, see Lenhard (2014) for a more detailed account of DFT and the development of computational chemistry.

Whereas the adjustment of parameters – to use the more neutral terminology – is pivotal for matching given data, i.e. for predictive success, this very success condition also entails a serious disadvantage.⁷ Complicated schemes of adjusted parameters might block theoretical progress. In our climate case, any new cloud parameterization that intends to work with a more thorough theoretical understanding has to be developed for many years and then has to compete with a well-tuned forerunner. Again, this kind of problem is more general. In quantum chemistry, many-parameter adaptations of density functionals have brought great predictive success but at the same time render the rational re-construction of why such success occurs hard, if not impossible (Perdew et al. 2005, discussed in Lenhard 2014). The situation in thermodynamics is similar, cf. Hasse and Lenhard (2017).

Let us take stock regarding the first argument for the erosion of modularity. Tuning, or adjusting, parameters is not merely an *ad hoc* procedure to smoothen a model, rather it is a pivotal component for simulation modeling. Tuning convolutes heterogeneous parts that do not have a common theoretical basis. Tuning proceeds holistically, on basis of global model behavior. How particular parts function often remains opaque. By interweaving local and global considerations, and by convoluting the interdependence of various parameter choices, tuning destructs modularity.

Looking back to Simon's clockmaker story, we see that its basic setting does not match the situation in a fundamental way. The perfect cogwheel picture is misleading, because it presupposes a clear identification of mechanisms and their interactions. In our examples, we saw

⁷ There are other dangers, like over-fitting, that I leave aside.

that building a simulation model, different from building a clockwork, cannot proceed top-down. Moreover, different modules and their interfaces get convoluted during the processes of mutual adaptation.

4. Erosion of modularity 2: kluging

The second argument for the erosion of modularity approaches the matter from a different angle, namely from a certain practice in developing software known as kluging (also spelled kludging)⁸. “Kluge” is a term from colloquial language that became a term in computer slang. I remember when back in my childhood our family and another, befriended one drove towards holidays in two cars. In the middle of the night, while crossing the Alps, the exhaust pipe of our friends before us broke, creating a shower of sparks where the pipe met the asphalt. There was no chance of getting the exhaust pipe repaired, but the father did not hesitate long and used his necktie to fix it provisionally.

The necktie worked as a kluge, which is in the words of Wikipedia “a workaround or quick-and-dirty solution that is clumsy, inelegant, difficult to extend and hard to maintain, yet an effective and quick solution to a problem.” The notion has been incorporated and become popular in the language of software programming and is closely related to the notion of bricolage.

Andy Clark, for instance, stresses the important role played by kluges in complex modular computer modeling. For him, a kluge is “an inelegant, ‘botched together’ piece of program; something functional but somehow messy and unsatisfying”, it is—Clark refers to Sloman—“a piece of program or machinery which works up to a point but is very complex, unprincipled in its design, ill-understood, hard to prove complete or sound and therefore having unknown limitations, and hard to maintain or extend”. (Clark 1987, 278)

Kluges carried forward their way from programmers’ colloquial language into the body of philosophy guided by scholars like Clark and Wimsatt who are inspired both by computer

⁸ Both spellings „kluge“ and „kludge“ are used. There is not even agreement of how to pronounce the word. In a way, that fits to the very concept. I will use “kluge“, but will not change the habits of other authors cited with “kludge“.

modeling and evolutionary theory.⁹ The important point in our present context is that kluges may function for a whole system, i.e. for the performance of the entire simulation model, whereas it has no meaning in relation to the submodels and modules: “what is a kludge considered as an item designed to fulfill a certain role in a large system, may be no kludge at all when viewed as an item designed to fulfill a somewhat different role in a smaller system.” (Clark 1987, 279)

Since kluging stems from colloquial language and is not seen as a good practice anyway, examples cannot be found easily in published scientific literature. This observation notwithstanding, kluging is a widely occurring phenomenon. Let me give an example that I know from visiting an engineering laboratory. There, researchers (chemical process engineers) are working with simulation models of an absorption column, the large steel structures in which reactions take place under controlled conditions. The scientific details do not matter here, since the point is that the engineers build their model on the basis of a couple of already existing modules, including proprietary software that they integrate into their simulation without having access to the code. Moreover, it is common knowledge in the community that this (unknown) code is of poor quality. Because of programming errors and because of ill-maintained interfaces, using this software package requires modifications on the part of the remaining code outside the package. These modifications are there for no good theoretical reason, albeit for good practical reasons. They make the overall simulation run as expected (in known cases); and they allow working with existing software. The modifications thus are typical kluges.

Again, kluging occurs in virtually every site where large software programs are built. Simulation models hence are a prime instance, especially when the modeling steps of one group build on the results (models, software packages) of other groups. One common phenomenon is the increasing importance of “exception handling”, i.e. of finding effective repairs when the software, or the model, performs in unanticipated and undesired ways. In this situation, the software might include a bug that is invisible (does not affect results) most of the time, but becomes effective under particular conditions. Often extensive testing is needed for finding out about unwanted behavior that occurs in rare and particular situations that are conceived of as “exceptions”, indicating that researchers do not aim at a major reconstruction, but at a local repair,

⁹ The cluster of notions like bricolage and kluging common in software programming and biological evolution would demand a separate investigation. See, as a teaser, Francois Jacob’s account of evolution as bricolage (1994).

counteracting this particular exception. Exception handling can be part of a sound design process, but increased use of exception handling is symptomatic of excessive kluging.

Presumably all readers who ever contributed to a large software program know about experiences of this kind. It is commonly accepted that the more comprehensive a piece of software gets, the more energy for exception handling new releases will require. Operating systems of computers, for example, often receive weekly patches. Many scientists who work with simulations are in a similar situation, though not obviously so.

If, for instance, meteorologists want to work on, say, hurricanes, they will likely take a meso-scale (multi-purpose) atmospheric model from the shelf of some trusted modeling center and add specifications and parameterizations relevant for hurricanes. Typically, they will not know in exactly what respects the model had been tuned, and also lack much other knowledge about strengths and weaknesses of this particular model. Consequently, when preparing their hurricane modules, they will add measures into their new modules that somehow balance out undesired model behavior. These measures can also be conceived as kluges.

Why should we see these examples as typical instances and not as exceptions? Because they arise from practical circumstances of developing software, which is a core part of simulation modeling. Software engineering is a field that was envisioned as the “professional” answer to the increasing complexity of software. And I frankly admit that there are well-articulated concepts that would in principle ensure software is clearly written, aptly modularized, well maintained, and superbly documented. However, the problem is that science *in principle* is different from science *in practice*.

In practice, there are strong and constant forces that drive software development into resorting to kluges. Economic considerations are always a reason, be it on the personal scale of research time, be it on the grand scale of assigning teams of developers to certain tasks. Usually, software is developed “on the move”, i.e. those who write it have to keep up with changing requirements and a narrow timeline, in science as well as industry. Of course, in the ideal case the implementation is tightly modularized. A virtue of modularity is that it is much quicker incorporating “foreign” modules than developing them from scratch.

If these modules have some deficiencies, however, the developers will usually not start a fundamental analysis of how unexpected deviations occurred, but rather spend their energy for

adapting the interfaces so that the joint model will work as anticipated in the given circumstances. In common language: repair, rather than replace. Examples reach from integrating a module of atmospheric chemistry into an existing general circulation model up to implementing the new version of the operating system of your computer. Working with complex computational and simulation models seems to require a certain division of labor and this division, in turn, thrives on software traveling easily. At the same time, this will provoke kluges on the side of those that try to connect software modules.

Kluges thus arise from unprincipled reasons: throw-away code, made for the moment, is not replaced later, but becomes forgotten, buried in more code, and eventually fixed. This will lead to a cascade of kluges. Once there, they prompt more kluges, tending to become layered and entrenched.¹⁰

Foote and Yoder, prominent leaders in the field of software development, give an ironic and funny account of how attempts to maintain a rationally designed software architecture constantly fail in practice.

“While much attention has been focused on high-level software architectural patterns, what is, in effect, the de-facto standard software architecture is seldom discussed. This paper examines this most frequently deployed of software architectures: the BIG BALL OF MUD. A big ball of mud is a casually, even haphazardly, structured system. Its organization, if one can call it that, is dictated more by expediency than design. Yet, its enduring popularity cannot merely be indicative of a general disregard for architecture. (...) 2. Reason for degeneration: ongoing evolutionary pressure, piecemeal growth: Even systems with well-defined architectures are prone to structural erosion. The relentless onslaught of changing requirements that any successful system attracts can gradually undermine its structure. Systems that were once tidy become overgrown as piecemeal growth gradually allows elements of the system to sprawl in an uncontrolled fashion.” (Foote and Yoder 1999, ch. 29)

I would like to repeat the statement from above that there is no necessity in the corruption of modularity and rational architecture. Again, this is a question of science in practice vs. science in principle. “A sustained commitment to refactoring can keep a system from subsiding into a big

¹⁰ Wimsatt (2007) writes about “generative entrenchment” when speaking about the analogy between software development and biological evolution, see also Lenhard and Winsberg (2010).

ball of mud,” Foote and Yoder concede. There are even directions in software engineering that try to counteract the degradation into Foote’s and Yoder’s big ball of mud. The movement of “clean code“, for instance, is directed against what Foote and Yoder describe. Robert Martin, the pioneer of this school, proposes to keep code clean in the sense of not letting the first kluge slip in. And surely there is no principled reason why one should not be able to avoid this. However, even Martin accepts the diagnosis of current practice.

Similarly, Richard Gabriel (1996), another guru of software engineering, makes the analogy to housing architecture and Alexander’s concept of “habitability”, which intends to integrate modularity and piecemeal growth into one “organic order”. Anyway, when describing the starting point, he more or less duplicates what we heard above from Foote and Yoder.

Finally, I want to point out that the matter of kluging is related to what is discussed in philosophy of science under the heading of opacity (like in Humphreys 2009). Highly kluged software becomes opaque. One can hardly disentangle the various reasons that led to particular pieces of code, because kluges are sensible only in the particular context at the time. In this important sense, simulation models are historical objects. They carry around – and depend on – their history of modifications. There are interesting analogies with biological evolution that have become a topic when complex systems had become a major issue in discussion computer use. Winograd and Flores, for instance, come to a conclusion that also holds in our context here: “each detail may be the result of an evolved compromise between many conflicting demands. At times, the only explanation for the system’s current form may be the appeal to this history of modification.“ (1991, 94)¹¹

Thus, the brief look into the somewhat elusive field of software development has shown us that two conditions foster kluging. First, the exchange of software parts that is more or less motivated by flexibility and economic requirements. This thrives on networked infrastructure. Second, iterations and modifications are easy and cheap. Due to the unprincipled nature of kluges, their construction requires repeated testing whether they actually work in the factual circumstances. Kluges hence fit to the exploratory and iterative mode of modeling that characterizes

¹¹ Interestingly, Jacob (1994) gives a very similar account of biological evolution when he writes that simpler objects are more dependent on (physical) constraints than on history, while history plays the greater part when complexity increases.

simulations. Furthermore, layered kluges solidify themselves. They make code hard or impossible to understand; modifying pieces that are individually hard to understand will normally lead to a new layer of kluges – and so on. Thus, kluging makes modularity erode and this is the second argument why simulation modeling systematically undermines modularity.

5. The limits of validation

What does the erosion of modularity mean for the validation of computer simulations? We have seen that the power and scope of simulation is built on the tendency toward holism. But holism and the erosion of modularity are two sides of the same coin. The key point regarding methodology is that holism is driven by the very procedure that makes simulation so widely applicable! It is through adjustable parameters that simulation models can be applied to systems beyond the control of theory (alone). It is through this very strategy that modularity erodes.

One ramification of utmost importance is about the concept of validation. In the context of simulation models the community speaks of verification and validation, or “V&V”. Both are related, but the unanimous advice in the literature is to keep them separate. While verification checks the model internally, i.e. whether the software indeed captures what it is supposed to, validation checks whether the model adequately represents the target system. A standard definition states that “verification [is] the process of determining that a model implementation accurately represents the developer’s conceptual description of the model and the solution to the model.” While validation is defined as “the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.” (Oberkampf and Trucano 2000, 3) Though there is some leeway of defining V&V, you get the gist of it from the saying: verification checks whether the model is right¹², while validation checks whether we have the right model.

Due to the increasing usage and growing complexity of simulations, the issue of V&V is itself a growing field in simulation literature. One example is the voluminous monograph by Oberkampf and Roy (2010) that meticulously defines and discusses the various steps to be included in V&V procedures. A first move in this analysis is to separate model form from model parameters. Each

¹² This sloppy saying should not obscure that the process of verification comprises a package of demanding tasks.

parameter then belongs to a particular type of parameter that determines which specific steps in V&V are required. Oberkampff gives the following list of model parameter types:

- “
- measurable properties of the system or the surroundings,
 - physical modeling parameters,
 - ad hoc model parameters,
 - numerical algorithm parameters,
 - decision parameters,
 - uncertainty modeling parameters.” (Oberkampff and Roy 2010, section 13.5.1, p.623)

My point is that the adjustable parameters we discussed are of a type that is evading the V&V fencing. These parameters cannot be kept separate from the model form, since the form alone does not capture representational (nor behavioral) adequacy. A cloud parameterization scheme makes sense only with parameter values already assigned and the same holds for a many-parameter density functional. Before the process of adjustment, the mere form of the functional does not offer anything to be called adequate or inadequate. In simulation models, as we have seen, (predictive) success and adaptation are entangled.

The separation of verification and validation thus cannot be fully maintained in practice. It is not possible to first verify that a simulation model is ‘right’ before tackling the ‘external’ question whether it is the right model. Performance tests hence become the main handle for confirmation. This is a version of confirmation holism that points toward the limits of analysis. This does not lead to a complete conceptual breakdown of verification and validation. Rather, holism comes in degrees¹³ and is a pernicious tendency that undermines the verification-validation divide.¹⁴

Finally, we come back to the analogy, or rather dis-analogy between computer and clockwork. In an important sense, computers are not amplifiers, i.e. they are not analogous to gigantic clockworks. They do not (simply) amplify the force of mathematical modeling that has got stuck

¹³ I thank Rob Muir for pointing this out to me.

¹⁴ My conclusion about the inseparability of verification and validation is in good agreement with Winsberg’s more specialized claim in (2010) where he argues about model versions that evolve due to changing parameterizations, which has been criticized by Morrison (2015). As far as I can see, her arguments do not apply to the case made in this paper, which rests on a tendency toward holism, rather than a complete conceptual breakdown.

in too demanding operations. Rather, computer simulation is profoundly *changing* the setting of how mathematics is used.

In the present paper I questioned the rational picture of design. Also Brooks did this when he observed that Pahl and Beitz had to include more and more steps to somehow capture an unwilling and complex practice of design, or when he refers to Donald Schön who criticized a one-sided “technical rationality” that underlies the Rational Model (Brooks 2010, chapter 2). However, my criticism works, if you want, from ‘within’. It is the very methodology of simulation modeling, and how it works in practice, that challenges the rational picture by making modularity erode.

The challenge to the rational picture has quite fundamental ramifications because this picture influenced so many ways we conceptualize our world. I will spare the philosophical discussion of how simulation modeling is challenging our concept of mathematization and with it our picture of scientific rationality for another paper. Just let me mention the philosophy of mind as one example. How we are inclined to think about mind today is deeply influenced by the computer and by our concept of mathematical modeling. Jerry Fodor has defended a most influential thesis that mind is composed of information-processing devices that operate largely separately (Fodor 1983). Consequently, re-thinking how computer models are related to modularity invites to re-thinking the computational theory of the mind.

I would like to thank ...

References

- Agre, Philip E., Hierarchy and History in Simon’s “Architecture of Complexity“, *Journal of the Learning Sciences*, 3, 2003, 413-426.
- Brooks, Frederic P, *The Design of Design*. Boston, MA: Addison-Wesley, 2010.
- Clark, Andy, The Kludge in the Machine, in: *Mind and Language* 2(4), 1987, 277-300.
- Fillion, Nicolas, 2017, The Vindication of Computer Simulations, in Lenhard, J., and Carrier, M. (eds.), *Mathematics as a Tool*, Boston Studies in History and Philosophy of Science, forthcoming.
- Fodor, Jerry: *The Modularity of Mind*, 1983, MIT Press, Cambridge, MA.
- Foote, Brian und Joseph Yoder, *Pattern Languages of Program Design 4 (= Software Patterns. 4)*. Addison Wesley, 1999.

- Frigg, Roman and Julian Reiss, The Philosophy of Simulation. Hot New Issues or Same Old Stew?, in: *Synthese*, 169(3), 593-613, 2009.
- Gabriel, Richard P.: *Patterns of Software. Tales From the Software Community*, New York and Oxford: Oxford University Press, 1996.
- Gramelsberger, Gabriele und Johann Feichter (eds.): *Climate Change and Policy. The Calculability of Climate Change and the Challenge of Uncertainty*, Heidelberg: Springer 2011.
- Hasse, Hans, and Lenhard, J. (2017), On the Role of Adjustable Parameters, in Lenhard, J., and Carrier, M. (eds.), *Mathematics as a Tool*, Boston Studies in History and Philosophy of Science, forthcoming.
- Humphreys, Paul, The Philosophical Novelty of Computer Simulation Methods, *Synthese*, 169 (3):615 - 626 (2009).
- Jacob, Francois, *The Possible and the Actual*, Seattle: University of Washington Press, 1994.
- Lenhard, Johannes, *Disciplines, Models, and Computers: The Path To Computational Quantum Chemistry*, Studies in History and Philosophy of Science Part A, 48 (2014), 89-96.
- Lenhard, Johannes and Eric Winsberg, *Holism, Entrenchment, and the Future of Climate Model Pluralism*, in: Studies in History and Philosophy of Modern Physics, 41, 2010, 253-262.
- Mauritsen, Thorsten, Bjorn Stevens, Erich Roeckner, Traute Crueger, Monika Esch, Marco Giorgetta, Helmuth Haak, Johann Jungclaus, Daniel Klocke, Daniela Matei, Uwe Mikolajewicz, Dirk Notz, Robert Pincus, Hauke Schmidt, and Lorenzo Tomassini, Tuning the climate of a global model, *Journal of Advances in Modeling Earth Systems*, 4, 2012.
- Morrison, Margaret, *Reconstructing Reality. Models, Mathematics, and Simulations*. New York: Oxford University Press, 2015.
- Oberkampff, William L., and Roy, Christopher J., *Verification and Validation in Scientific Computing*. Cambridge, MA: Cambridge University Press, 2010.
- Oberkampff, William L. and Trucano, T.G., *Validation Methodology in Computational Fluid Dynamics*, American Institute for Aeronautics and Astronautics, 2000 – 2549, 2000.
- Pahl, G. and Beitz, W. 1984. *Engineering Design: A Systematic Approach*. Revised editions in 1996, 2007. Berlin: Springer.
- Parker, Wendy, Values and Uncertainties in Climate Prediction, revisited, *Studies in History and Philosophy of Science* 2013.
- Perdew, J. P., Ruzsinsky, A., Tao, J., Staroverov, V., Scuseria, G., & Csonka, G. (2005). Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *The Journal of Chemical Physics*, 123.
- Royce, Wynston, Managing the Development of Large software Systems. *Proceedings of IEEE WESCON* 26 (August), 1970, 1–9.
- Simon, Herbert A., *The Sciences of the Artificial*, Cambridge, MA: The MIT Press, 1969.
- Smith, Leonard A., What Might We learn From Climate Forecasts?, in: *Proceedings of the National Academy of Sciences USA*, 4(99), 2002, 2487-2492.
- Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.), *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel*

on Climate Change, 2007. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Stainforth, D.A., Downing, T.E., Washington, R. and New, M. (2007) Issues in the interpretation of climate model ensembles to inform decisions, *Philosophical Transactions of the Royal Society*, Volume 365, Number 1857, 2145-2161.

Wimsatt, William C., *Re-Engineering Philosophy for Limited Beings. Piecewise approximations to reality*, Cambridge, MA and London, England: Harvard University Press, 2007.

Winograd, Terry und F. Flores, *Understanding Computers and Cognition*, Reading, MA: Addison-Wesley, 1991.

Winsberg, Eric, *Science in the Age of Computer Simulation*, Chicago, Ill.: University of Chicago Press, 2010.

Accuracy, conditionalization, and probabilism

Peter J. Lewis, University of Miami

Don Fallis, University of Arizona

March 3, 2016

Abstract

Accuracy-based arguments for conditionalization and probabilism appear to have a significant advantage over their Dutch Book rivals. They rely only on the plausible epistemic norm that one should try to decrease the inaccuracy of one's beliefs. Furthermore, it seems that conditionalization and probabilism follow from a wide range of measures of inaccuracy. However, we argue that among the measures in the literature, there are some from which one can prove conditionalization, others from which one can prove probabilism, and none from which one can prove both. Hence at present, the accuracy-based approach cannot underwrite both conditionalization and probabilism.

A central concern of epistemology is uncovering the rational constraints on an agent's credences, both at a time and over time. At a time, it is typically maintained that an agent's credences should conform to the probability axioms, and over time, it is often maintained that an agent's credences should conform to conditionalization. How could such norms be justified? The traditional approach is to show that if your credences violate these norms, then there is a set of bets, each of which you consider fair, but which collectively are such that if you accept them all you will lose money whatever happens. Since you do not want to be a "money pump", you should adopt coherent credences. However, this *Dutch book* strategy rests on controversial assumptions concerning prudential rationality and its connection to epistemic rationality.

The prudential elements may not be essential to the Dutch book approach (Vineberg 2012). But even so, it would be better to be able to derive probabilism and conditionalization from a clearly epistemic basic norm. A more

recent approach seeks to do precisely that: to derive probabilism and conditionalization from the intuitive epistemic norm that you should endeavor to make your credences as accurate—as close to the truth—as possible. Drawing on the work of Joyce (1998; 2009), Greaves and Wallace (2006) and Predd et al. (2009), Pettigrew (2013) argues that the accuracy-based approach vindicates both probabilism and conditionalization. We argue that this conclusion is too strong: at present, the accuracy-based approach can vindicate *either* conditionalization *or* probabilism, but not both.

Our argument turns on the features of various proposed measures of accuracy. The accuracy-based approach is predicated on the assumption that the accuracy of your credences can be measured. Pettigrew (2013, 905) argues that it is a strength of the accuracy-based approach that conditionalization and probabilism follow from a wide range of measures, so that it doesn't matter which measure is used to assess the accuracy of an agent's credences. Our counter-argument is that it does matter: of the known measures, some vindicate conditionalization, and some vindicate probabilism, but there is no known measure of inaccuracy from which both conditionalization and probabilism can be derived.

1 Accuracy and conditionalization

First, let us briefly run through the argument via which conditionalization and probabilism are claimed to follow from considerations of accuracy, starting with conditionalization. Suppose you have credences $\mathbf{b} = (b_1, b_2, \dots, b_n)$ in propositions $\mathbf{X} = (X_1, X_2, \dots, X_n)$, where the propositions form a partition, i.e. they are exhaustive and mutually exclusive, so that exactly one of them is true. The accuracy approach takes it that your primary epistemic goal is having credences that are as accurate as possible, where complete accuracy is a credence of 1 in the true proposition and a credence of 0 in each of the false propositions. The closer your credences are to complete accuracy, the better.

For this epistemic goal to make sense, we need a measure of closeness. In what follows we will discuss several such measures, expressed as measures of *inaccuracy*: the larger the measure, the further your credences are from the truth. Hence your goal is to minimize the value of this inaccuracy measure. By far the dominant measure in the literature is the quadratic rule or Brier rule, which takes the square of the difference between your credence in each

proposition and its truth value, and sums the results. So for a partition, if $I_i(\mathbf{b})$ is the inaccuracy of credences \mathbf{b} when proposition X_i is true, then the Brier rule can be expressed as follows:¹

Simple Brier rule: $I_i(\mathbf{b}) = (1 - b_i)^2 + \sum_{j \neq i} b_j^2$.

The Brier rule has been defended by epistemologists (Joyce 2009, 290; Leitgeb and Pettigrew 2010, 219), and is frequently cited as the prime example of an inaccuracy measure (Greaves and Wallace 2006, 627; Pettigrew 2013, 899).

Suppose you obtain evidence E that is consistent with some but not all of the propositions \mathbf{X} . How should you distribute your credence over the remaining propositions? If your goal is to minimize your inaccuracy, presumably the best you can do is to minimize your *expected* inaccuracy given your prior credences \mathbf{b} . So suppose that after you learn E , you shift your credence in proposition X_i from b_i to x . If X_i is true, the contribution of this new credence to your overall inaccuracy is $(1 - x)^2$, and if X_i is false, the contribution is x^2 . Given your prior credences \mathbf{b} , you judge that the chance that X_i is true is b_i , and the chance that X_i is false is $\sum_{E-i} b_j$, where the notation $E - i$ indicates that the sum is over all propositions consistent with E except X_i . That is, the total contribution C of this new credence to your expected inaccuracy is given by:

$$C = (1 - x)^2 b_i + x^2 \sum_{E-i} b_j.$$

Your goal is to minimize C . So consider where $dC/dx = 0$:

$$\begin{aligned} \frac{dC}{dx} &= -2(1 - x)b_i + 2x \sum_{E-i} b_j \\ &= -2b_i + 2x \sum_E b_j, \end{aligned}$$

where the sum in the last line is now over all propositions consistent with E . This expression is zero when

$$x = \frac{b_i}{\sum_E b_j}.$$

¹We call the version of the Brier rule applicable to a partition the *simple* Brier rule only for ease of reference (and similarly for the simple log rule and simple spherical rule to be introduced later).

But note that this value for x is just your prior credence in X_i conditional on E :

$$c(X_i|E) = \frac{c(X_i \wedge E)}{c(E)} = \frac{b_i}{\sum_E b_j}.$$

That is, conditionalizing on E minimizes your expected inaccuracy.² So if your epistemic goal is to minimize inaccuracy, you should conditionalize on new evidence.

Greaves and Wallace (2006) generalize this proof to cover measures of inaccuracy other than the Brier rule. In particular, they show that conditionalization minimizes expected inaccuracy for any measure of inaccuracy $I_i(\mathbf{b})$ satisfying *strict propriety*:

Strict propriety: For any distinct probabilistic credences \mathbf{b} and \mathbf{b}' , $\sum_i b_i I_i(\mathbf{b}) < \sum_i b_i I_i(\mathbf{b}')$.

Strict propriety says that the expected inaccuracy of your current credences \mathbf{b} is lower than the expected inaccuracy of any alternative credences \mathbf{b}' you might adopt, where the expectation is calculated according to your current credences. If it fails, then the injunction to minimize inaccuracy makes your beliefs pathologically unstable: you can lower your expected inaccuracy by shifting your credences, even in the absence of new evidence. Hence strict propriety serves as a reasonable constraint on measures of inaccuracy. The Brier rule is strictly proper, as are several other proposed inaccuracy measures to be discussed below.

Greaves and Wallace begin by introducing some terminology. They say that a set of credences \mathbf{b} *recommends* a set of credences \mathbf{b}' iff the expected inaccuracy of \mathbf{b}' is at least as low as the expected inaccuracy of \mathbf{b} , where the expectation is calculated using credences \mathbf{b} :

Recommendation: \mathbf{b} recommends \mathbf{b}' iff $\sum_i b_i I_i(\mathbf{b}) \geq \sum_i b_i I_i(\mathbf{b}')$

Note that if the inaccuracy measure $I_i(\mathbf{b})$ satisfies strict propriety, then \mathbf{b} only recommends itself.

They further define *quasi-conditionalization* as a belief updating rule that stipulates that your credences on learning E should be some set *recommended* by your prior credences conditional on E . They then prove

²This proof is a simplified version of the one in Leitgeb and Pettigrew (2010).

that quasi-conditionalization is always optimal: whatever measure of inaccuracy you choose, strictly proper or not, the expected inaccuracy of quasi-conditionalizing is at least as low as the expected inaccuracy of any other updating rule. Then if your measure of inaccuracy is strictly proper, conditionalization itself is optimal, since for strictly proper measures, credences only recommend themselves. In fact, since the inequality in strict propriety is *strict*, conditionalization is strictly better than any other updating rule: it uniquely minimizes expected inaccuracy. As Pettigrew (2013, 905) notes, this is a strong result: any inaccuracy measure satisfying strict propriety can be used to vindicate conditionalization, and strict propriety is a constraint we would expect any reasonable inaccuracy measure to obey anyway.

2 Accuracy and probabilism

Now let us turn to the arguments that your credences at a time should obey the probability axioms. So far, we have been assuming that the propositions we are interested in form a partition. But the probability axioms include constraints on your credences in disjunctions, and to model such constraint we need to allow that more than one of the propositions you are considering can be true. To that end, suppose that you have credences $\mathbf{b} = (b_1, b_2, \dots, b_n)$ in propositions $\mathbf{X} = (X_1, X_2, \dots, X_n)$, where now the set of propositions forms a Boolean algebra, i.e. it is closed under negation and disjunction. So now we can no longer model a possible world simply as an index (picking out the unique true proposition); instead, we need to label each proposition separately as either true or false. That is, a possible world is specified by $\omega = (\omega_1, \omega_2, \dots, \omega_n)$, where $\omega_i = 1$ when X_i is true and $\omega_i = 0$ when X_i is false. In this context, the Brier rule can be rewritten as follows:

Symmetric Brier rule: $I(\omega, \mathbf{b}) = \sum_i (b_i - \omega_i)^2$.

As before, the inaccuracy of your beliefs according to the Brier rule is given by the sum of the squares of the distance of each belief from the relevant truth value. That is, the Brier rule is *symmetric*, in the sense that distance from the truth for a true proposition plays the same role as distance from falsity plays for a false proposition. This property will be important later.

The general strategy for defending probabilism based on accuracy goes as follows. Suppose that your current credences are incoherent—that is, they

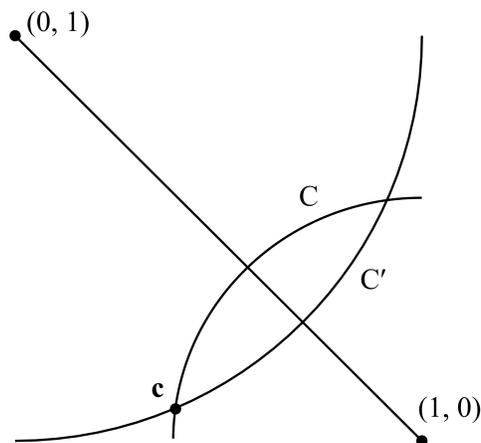


Figure 1: De Finetti's construction for a two-element partition (Joyce 1998, 582).

violate the probability axioms. Then one can appeal to a measure of inaccuracy to show that there are coherent credences that *dominate* your current credences—that are more accurate than your current credences whatever the truth values of the propositions concerned. If your goal is to minimize inaccuracy, this gives you a clear reason to avoid incoherent credences: there are always coherent credences that are more accurate, whatever the world is like.

De Finetti (1974, 87) constructs a dominance argument of this kind based on the Brier rule.³ For illustration, consider the simple case of a proposition and its negation: that is, the propositions under consideration are just $(X, \neg X)$. In this case the space of possible credences forms a plane, as shown in figure 1: your credence in X is the horizontal coordinate, and your credence in $\neg X$ is the vertical coordinate. The two possible worlds are represented by the points $(1, 0)$ and $(0, 1)$, and your credences obey the probability axioms if and only if they lie on the straight line that connects these two points, since along this line your credences in X and $\neg X$ sum to 1.

Suppose that your credences are incoherent: they are represented by a point $\mathbf{c} = (c_1, c_2)$ that lies *off* this diagonal. And suppose first that the

³As Joyce (1998, 580) notes, de Finetti sets up this argument in terms of bets. However, as Pettigrew (2013, 901) points out, it can be redescribed as an accuracy-based argument.

actual world is represented by the bottom-right corner $(1, 0)$ —i.e. X is true and $\neg X$ is false. Then the inaccuracy of your credences according to the Brier rule is $I(\omega, \mathbf{c}) = (1 - c_1)^2 + (c_2)^2$. Note that this is just the square of the Euclidean distance between (c_1, c_2) and $(1, 0)$. That is, every point on the circle segment C has the same inaccuracy as \mathbf{c} , and every point between C and $(1, 0)$ has a lower inaccuracy. Now suppose instead that the actual world is represented by the top-left corner $(0, 1)$ —i.e. X is false and $\neg X$ is true. Then the inaccuracy of your credences is $I(\omega, \mathbf{c}) = (c_1)^2 + (1 - c_2)^2$ —the square of the Euclidean distance between (c_1, c_2) and $(0, 1)$. That is, every point on the circle segment C' has the same inaccuracy as \mathbf{c} , and every point between C' and $(0, 1)$ has a lower inaccuracy.

Consider the area enclosed by the circle segments C and C' . The credences represented by the points in this area have a lower inaccuracy than \mathbf{c} if X is true and $\neg X$ false, and a lower inaccuracy than \mathbf{c} if X is false and $\neg X$ true. That is, they have a lower inaccuracy whatever the world is like. And this area includes part of the diagonal that represents coherent credences. So for any incoherent set of credences, there is a coherent set that is less inaccurate whatever the world is like. In this simple case, accuracy gives you a motive to adopt coherent credences.

In the general case, the space of possible credences is n -dimensional, where there are n propositions in the Boolean algebra. Each possible assignment of truth values to the n propositions is represented by a point in this space, and the set of coherent credences consists of these points plus the points on the straight lines that connect them, the points on the straight lines that connect those latter points, and so on. This set is called the *convex hull* V^+ of the possible truth value assignments V . Via a generalization of the construction of figure 1, de Finetti shows that if your credences are represented by a point that lies outside V^+ , then there are points in V^+ that are more accurate (according to the Brier rule) whichever point in the space represents the actual truth values of the propositions. Hence if you have incoherent credences, there are always coherent credences with a lower inaccuracy as measured by the Brier rule.

Predd et al. (2009) generalize this proof strategy to cover a wider range of inaccuracy measures. Their proof relies on two assumptions. The first is additivity:

Additivity: $I(\omega, \mathbf{b})$ can be expressed as $\sum_i s(\omega_i, b_i)$, where s is a continuous function of your credence in proposition X_i and its truth value.

Additivity states that the inaccuracy of your beliefs in a set of propositions is just the sum of your inaccuracies in the propositions taken individually—that is, $s(\omega_i, b_i)$ is the inaccuracy of your belief in proposition X_i , and $I(\boldsymbol{\omega}, \mathbf{b})$ is just the sum of these inaccuracies for all the propositions you are considering. Note that it also contains the requirement that the inaccuracy measure should be continuous. The Brier rule is obviously additive, since it is expressed as a sum over propositions.

The second assumption is a version of strict propriety. For an additive inaccuracy measure, strict propriety can be expressed in terms of your inaccuracy function for a single proposition $s(b_i, \omega_i)$ as follows:

Strict propriety (for an additive measure): $b_i s(x, 1) + (1 - b_i) s(x, 0)$ is uniquely minimized at $x = b_i$.

Predd et al. (2009) prove that any additive, strictly proper inaccuracy measure entails probabilism. De Finetti’s construction appeals to the natural distance measure implicit in the Brier rule—the Euclidean distance between two points in the space of your possible credences. But in the current case we have no explicit measure of inaccuracy, so Predd et al. appeal to a generalized “distance” measure⁴ called the Bregman divergence, defined for a strictly convex function $\Phi(\mathbf{x})$ as $d_\Phi(\mathbf{y}, \mathbf{x}) = \Phi(\mathbf{y}) - \Phi(\mathbf{x}) - \nabla\Phi(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})$. They show that if the inaccuracy measure $s(b_i, \omega_i)$ for a single proposition X_i is strictly proper, then the function $\varphi(b_i) = -b_i s(b_i, 1) - (1 - b_i) s(b_i, 0)$ is strictly convex. In terms of this function, Predd et al. show that for any additive, strictly proper inaccuracy measure, $I(\boldsymbol{\omega}, \mathbf{b}) = d_\Phi(\boldsymbol{\omega}, \mathbf{b})$, where $\Phi(\boldsymbol{\omega}) = \sum_i \varphi(\omega_i)$ and $\Phi(\mathbf{b}) = \sum_i \varphi(b_i)$.

The set of coherent credences forms a closed, convex subspace V^+ of the space of all possible credences. It is a fact from the theory of Bregman divergences that for any point \mathbf{c} outside V^+ , there is a unique point \mathbf{c}^* in V^+ such that $d_\Phi(\mathbf{c}^*, \mathbf{c}) \leq d_\Phi(\mathbf{y}, \mathbf{c})$ for all \mathbf{y} in V^+ . That is, \mathbf{c}^* is the unique closest point in V^+ to \mathbf{c} , using the Bregman divergence as a distance measure. It is a further fact that $d_\Phi(\mathbf{y}, \mathbf{c}^*) \leq d_\Phi(\mathbf{y}, \mathbf{c}) - d_\Phi(\mathbf{c}^*, \mathbf{c})$ for all \mathbf{y} in V^+ and \mathbf{c} outside V^+ . Note in particular that V^+ contains every possible world $\boldsymbol{\omega}$, since a consistent truth value assignment is also a coherent set of credences. So setting $\mathbf{y} = \boldsymbol{\omega}$, we have $d_\Phi(\boldsymbol{\omega}, \mathbf{c}^*) \leq d_\Phi(\boldsymbol{\omega}, \mathbf{c}) - d_\Phi(\mathbf{c}^*, \mathbf{c})$. Since d_Φ is a positive-valued function, $d_\Phi(\mathbf{c}^*, \mathbf{c}) > 0$, so $d_\Phi(\boldsymbol{\omega}, \mathbf{c}^*) < d_\Phi(\boldsymbol{\omega}, \mathbf{c})$, and hence

⁴The reason for the scare quotes is that the Bregman divergence is not symmetric, and distance measures are typically symmetric.

$I(\omega, \mathbf{c}^*) < I(\omega, \mathbf{c})$. That is, for any incoherent set of credences \mathbf{c} , there is a coherent set \mathbf{c}^* that is less inaccurate than \mathbf{c} in every possible world.

As Pettigrew (2013, 905) notes, this is a strong result: any inaccuracy measure satisfying strict propriety and additivity can be used to vindicate probabilism, and while additivity is perhaps not forced on us in the way that strict propriety is, it is certainly intuitive. As we shall see, there are several available measures satisfying additivity and strict propriety, so it initially looks like the accuracy-based program can justify both probabilism and conditionalization based on minimal premises. Our purpose in this paper is to argue that matters are not so straightforward.

3 Measures of inaccuracy

Let us return to the argument for conditionalization. This argument restricts inaccuracy measures to those that are strictly proper. Note that strict propriety is only a condition on *expected* inaccuracy. But expected inaccuracy is calculated on the basis of the *actual* inaccuracy that the measure in question ascribes to credences, and presumably there are a number of constraints any such measure must obey if it is to genuinely measure epistemic inaccuracy rather than something else. For example, if one of your credences shifts towards the truth, while your other credences stay the same, then clearly your actual inaccuracy should decrease. We wish to focus on one such constraint.

The constraint can be motivated by thinking about *elimination cases*. Suppose you are considering a set of mutually exclusive and exhaustive propositions, and suppose that your credences are coherent and that you conditionalize on evidence. You acquire some evidence that eliminates one false proposition—your credence in it becomes zero—but is uninformative regarding the other hypotheses—your credences in them remain in the same proportions. How does this affect the accuracy of your credences?

It seems obvious that your beliefs have become more accurate. If you believe that Tom, Dick or Harry might be the murderer (when in fact Tom did it), and you eliminate Harry while learning nothing about Tom or Dick, then you have made epistemic progress towards the truth, or at least away from falsity. It is true that your credence in the false proposition “Dick did it” goes up, but only by the same proportion that your credence in the true proposition “Tom did it” goes up.

Unfortunately, the simple Brier rule does not always concur. Let X_1 be

“Tom did it”, X_2 be “Dick did it”, and X_3 be “Harry did it”, where unknown to you X_1 is true. Suppose that your initial credences in (X_1, X_2, X_3) are $\mathbf{b} = (1/7, 3/7, 3/7)$. Then according to the simple Brier rule, your initial inaccuracy is $54/49 = 1.10$. Now suppose you acquire some evidence that eliminates X_3 , but is uninformative regarding X_1 and X_2 . That is, your credence in X_3 becomes 0 and your credences in X_1 and X_2 stay in the same proportions, so that your final credences are $\mathbf{b}^* = (1/4, 3/4, 0)$. Then according to the simple Brier rule, your final inaccuracy is $18/16 = 1.13$. That is, the Brier rule erroneously says that the inaccuracy of your beliefs has gone up.

For a measure to genuinely measure the actual inaccuracy of your beliefs, it should not be susceptible to counterexamples of this kind; it should count elimination cases as epistemically positive. That is, measures of inaccuracy should obey the following principle:

M: For coherent credences over a partition, if \mathbf{b} assigns a zero credence to some false proposition to which \mathbf{b}' assigns a non-zero credence, and credences in the remaining propositions stay in the same ratios, then \mathbf{b} is epistemically better than \mathbf{b}' .

The simple Brier rule, as the example shows, violates M, and hence does not plausibly measure the actual inaccuracy of your beliefs.⁵

Fortunately, though, there are alternative inaccuracy measures for partitions we can appeal to. The two most frequently mentioned are the simple log rule and the simple spherical rule:

Simple log rule: $I_i(\mathbf{b}) = -\ln b_i$

Simple spherical rule: $I_i(\mathbf{b}) = 1 - b_i / \sqrt{\sum_j b_j^2}$.

As before, $I_i(\mathbf{b})$ is the inaccuracy of credences \mathbf{b} when proposition X_i is true. Both of these measures satisfy M, and hence are not susceptible to elimination counterexamples.⁶ Hence each can plausibly be claimed to measure epistemic inaccuracy. Furthermore, each is strictly proper, and so each can be used to

⁵One might reasonably think that acceptable measures of accuracy should obey a stronger principle than M; see (*reference removed*).

⁶This is trivial for the log rule, and easily proven for the spherical rule. See (*reference removed*).

underwrite conditionalization via the above argument strategy. So there are some inaccuracy measures that vindicate conditionalization, but not all strictly proper measures do so. In particular, the simple Brier rule cannot be used to vindicate conditionalization.

But what about probabilism? The simple log rule and simple spherical rule are not applicable to a Boolean algebra, and so cannot be used to prove probabilism as they stand. Perhaps the most straightforward way to generalize them is simply to sum the contribution given by the simple rule for each true proposition in the Boolean algebra, while ignoring the false propositions in the algebra:

Asymmetric log rule: $I(\mathbf{b}, \omega) = \sum_i F(\omega_i, b_i)$, where $F(0, b_i) = 0$ and $F(1, b_i) = -\ln b_i$.

Asymmetric spherical rule: $I(\mathbf{b}, \omega) = \sum_i F(\omega_i, b_i)$, where $F(0, b_i) = 0$ and $F(1, b_i) = 1 - b_i / \sqrt{\sum_j b_j^2}$.

Both these rules are asymmetric, in the sense that inaccuracy is calculated differently for true and false propositions. These rules satisfy principle M: for coherent credences, if your credence in a false proposition goes down and your remaining credences stay in the same ratios, then your credence in each true proposition goes up, and so your inaccuracy according to the relevant asymmetric rules goes down. Hence the asymmetric log and spherical rules are immune from elimination counterexamples.

But these rules do not satisfy the combination of additivity and strict propriety required for the proof of probabilism. The asymmetric spherical rule is not additive: $F(1, b_i)$ is not a function of b_i alone. The asymmetric log rule is additive, but it is not strictly proper in the required sense: $F(1, b_i)$ is strictly proper, but $F(0, b_i)$ is not. Indeed, it is straightforward to show directly that these rules cannot be used as the basis of a dominance argument for probabilism. Consider, for example, a two element partition, and the incoherent credence assignment $(1, 1)$. The asymmetric log rule counts these incoherent credences as *perfectly* accurate (since the credence in the false proposition is ignored), so no coherent credences can dominate them. According to the asymmetric spherical rule, multiplying all credences by a constant has no effect on inaccuracy, so this assignment has the same inaccuracy as the coherent credence assignment $(1/2, 1/2)$. If coherent assignments cannot be dominated, then neither can the initial incoherent assignment.

But if coherent assignments *can* be dominated then the dominance proof of probabilism fails anyway.

So the asymmetric versions of the log rule and the spherical rule cannot be used to prove probabilism. But for a Boolean algebra, the log rule and the spherical rule are usually given a formulation that is symmetric between truth and falsity:

Symmetric log rule: $I(\omega, \mathbf{b}) = \sum_i -\ln |(1 - \omega_i) - b_i|$

Symmetric spherical rule: $I(\omega, \mathbf{b}) = \sum_i 1 - \frac{|(1 - \omega_i) - b_i|}{\sqrt{b_i^2 + (1 - b_i)^2}}$

(see e.g. Joyce 2009, 275). These measures are additive, and each term in the sum is individually strictly proper, so they can each be used to prove probabilism via the proof of Predd et al.

But unfortunately, in their symmetric forms all three rules—Brier, log and spherical—are subject to elimination counterexamples. For the Brier rule, the counterexample is the same as before, since the symmetric Brier rule reduces to the simple Brier rule when applied to a partition.⁷ That is, consider a credence shift from $\mathbf{b} = (1/7, 3/7, 3/7)$ to $\mathbf{b}^* = (1/4, 3/4, 0)$ when X_1 is true. According to the symmetric Brier rule, your initial inaccuracy is 1.10, and your final inaccuracy is 1.13, so your inaccuracy goes up. And this example works equally well against the symmetric spherical rule: according to this rule, your initial inaccuracy is 1.24 and your final inaccuracy is 1.37, so your inaccuracy goes up. This particular counterexample does not work against the symmetric log rule, but a similar one does. Suppose your initial credences are $\mathbf{b} = (1/13, 6/13, 6/13)$, and your final credences are $\mathbf{b}^* = (1/7, 6/7, 0)$. Then according to the symmetric log rule your initial inaccuracy is 3.80, and your final inaccuracy is 3.89: your inaccuracy goes up. Hence the symmetric measures all violate principle M, and so none of them can be used to prove conditionalization.

⁷Strictly, applying these rules to a Boolean algebra requires including credences in the negations $\neg X_1$, $\neg X_2$ and $\neg X_3$, plus the tautology $X_1 \vee X_2 \vee X_3$ and the contradiction $\neg(X_1 \vee X_2 \vee X_3)$. But for coherent credences the inaccuracies of the tautology and the contradiction are zero, and for symmetric rules the inaccuracy of $\neg X_i$ is the same as that of X_i , so the inaccuracy calculated over the entire Boolean algebra is simply twice the inaccuracy over the partition (X_1, X_2, X_3) .

4 The extent of the problem

Let us sum up. The simple Brier rule cannot be used to prove conditionalization, but the simple log and spherical rules can. The obvious generalizations of the simple log and spherical rules to a Boolean algebra—the asymmetric log and spherical rules—cannot be used to prove probabilism. The symmetric Brier, log and spherical rules can be used to prove probabilism, but none of them underwrites conditionalization. So we have found no measure that can be used to prove both conditionalization *and* probabilism.

Could there be such a measure? Perhaps, although it is worth noting that one can prove that *any* inaccuracy measure that satisfies additivity, strict propriety and a plausible symmetry principle is subject to elimination counterexamples. The symmetry principle is precisely the one discussed above—that the inaccuracy measure treats truth the same as falsity, in the sense that it is a function of the distance between each credence and its respective truth value. For an additive inaccuracy measure, the symmetry principle can be expressed in terms of the inaccuracy function for a single proposition $s(\omega_i, b_i)$ as follows:

Symmetry: $s(\omega_i, b_i) = s(|1 - \omega_i|, |1 - b_i|)$.

It is certainly highly plausible that this is part of what it means for s to measure your distance from the truth, and as discussed above, the typical Boolean algebra forms of the Brier rule, log rule and spherical rule all satisfy it.

Let us see how this symmetry principle, together with additivity and strict propriety, lead to elimination counterexamples. Consider a single proposition X_i in which your credence is $b_i = 1/2$. According to strict propriety, the quantity $(1/2)s(1, x) + (1/2)s(0, x)$ must be uniquely minimized at $x = 1/2$. In particular, the value of this expression for $x = 1/2$ must be lower than its value for $x = 1$:

$$(1/2)s(1, 1/2) + (1/2)s(0, 1/2) < (1/2)s(1, 1) + (1/2)s(0, 1),$$

and for $x = 0$:

$$(1/2)s(1, 1/2) + (1/2)s(0, 1/2) < (1/2)s(1, 0) + (1/2)s(0, 0).$$

Adding these:

$$s(1, 1/2) + s(0, 1/2) < (1/2)s(1, 1) + (1/2)s(0, 1) + (1/2)s(1, 0) + (1/2)s(0, 0).$$

But by symmetry, $s(1, 1/2) = s(0, 1/2)$, $s(1, 1) = s(0, 0)$ and $s(0, 1) = s(1, 0)$. Substituting:

$$2s(0, 1/2) < s(0, 1) + s(0, 0).$$

Now consider your credences in three exhaustive and mutually exclusive propositions $\mathbf{X} = (X_1, X_2, X_3)$. Consider in particular the credence shift from $\mathbf{m} = (0, 1/2, 1/2)$ to $\mathbf{b} = (0, 1, 0)$ for truth values $\boldsymbol{\omega} = (1, 0, 0)$. By separability, $I(\boldsymbol{\omega}, \mathbf{m}) = s(1, 0) + 2s(0, 1/2)$, and $I(\boldsymbol{\omega}, \mathbf{b}) = s(1, 0) + s(0, 1) + s(0, 0)$. So since $2s(0, 1/2) < s(0, 1) + s(0, 0)$ it follows that $I(\boldsymbol{\omega}, \mathbf{m}) < I(\boldsymbol{\omega}, \mathbf{b})$: your inaccuracy goes up. But the shift from $\mathbf{m} = (0, 1/2, 1/2)$ to $\mathbf{b} = (0, 1, 0)$ is an elimination case: a false proposition is eliminated, and your credences in the remaining hypotheses stay in the same proportions. And lest one worry about the fact that your initial credence in the true proposition is zero, we can modify the example. Consider the credence assignments $\mathbf{m}' = (\delta/(2 + \delta), 1/(2 + \delta), 1/(2 + \delta))$ and $\mathbf{b}' = (\delta/(1 + \delta), 1/(1 + \delta), 0)$. For small δ these are close to \mathbf{m} and \mathbf{b} , and hence by the continuity clause of additivity, the inaccuracy of \mathbf{m}' remains lower than that of \mathbf{b}' . Again, the transition from \mathbf{m}' to \mathbf{b}' is an elimination case, and now your credence in the true proposition is non-zero.

So elimination counterexamples afflict any inaccuracy measure that satisfies additivity, strict propriety and symmetry. That is, any symmetric measure that satisfies the assumptions of Predd et al.'s proof of probabilism violates principle M, and hence cannot be used to prove conditionalization. Symmetry is not a premise in the Predd argument, so it is possible that an asymmetric measure might allow the derivation of both probabilism and conditionalization. But the only plausible asymmetric measure in the literature is the log rule (Bernardo 1979), and we have seen that the asymmetric log rule does not vindicate probabilism.

5 Conclusion

Pettigrew notes that conditionalization and probabilism follow from a wide range of measures of inaccuracy, and the implication is that it doesn't much matter which measure you pick. But we think it does matter. There are measures that vindicate conditionalization, and there are measures that vindicate probabilism, but nobody has yet identified a measure that vindicates both. Hence the accuracy-based approach does not, as yet, give us the justification we might want for the constraints on our credences.

References

- Bernardo, José M. (1979), “Expected information as expected utility”, *Annals of Statistics* 7: 686-690.
- de Finetti, Bruno (1974), *Theory of Probability*, vol. 1. New York: John Wiley and Sons.
- Greaves, Hilary and David Wallace (2006), “Justifying conditionalization: conditionalization maximizes expected epistemic utility”, *Mind* 115: 607–32.
- Joyce, James M. (1998), “A nonpragmatic vindication of probabilism”, *Philosophy of Science*, 65: 575–603.
- Joyce, James M. (2009), “Accuracy and coherence: prospects for an alethic epistemology of partial belief”, in F. Huber and C. Schmidt-Petri (eds.), *Degrees of Belief*. Dordrecht: Springer: 263–97.
- Leitgeb, Hannes, and Richard Pettigrew (2010), “An objective justification of Bayesianism I: measuring inaccuracy”, *Philosophy of Science* 77: 201–35.
- Pettigrew, Richard (2013), “Epistemic utility and norms for credence”, *Philosophy Compass* 8: 897–908.
- Predd, Joel B., Robert Seiringer, Elliott H. Lieb, Daniel N. Osherson, H. Vincent Poor, and Sanjeev R. Kulkarni (2009), “Probabilistic coherence and proper scoring rules”, *IEEE Transactions on Information Theory* 55: 4786–4792.
- Vineberg, Susan (2012), “Dutch book arguments”, in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/archives/spr2012/entries/dutch-book/>

Can Typicality Arguments Dissolve Cosmology's Flatness Problem?

C.D. McCoy*

20 February 2016

Abstract

The flatness problem in cosmology draws attention to a surprising fine-tuning of the spatial geometry of our universe towards flatness. Several physicists, among them Hawking, Page, Coule, and Carroll, have argued against the probabilistic intuitions underlying such fine-tuning arguments in cosmology and instead propose that the canonical measure on the phase space of Friedman-Robertson-Walker spacetimes should be used to evaluate fine-tuning. They claim that flat spacetimes in this set are actually typical on this natural measure and that therefore the flatness problem is illusory. I argue that they misinterpret typicality in this phase space and, moreover, that no conclusion can be drawn at all about the flatness problem by using the canonical measure alone.

For several decades now cosmologists have maintained that the old standard model of cosmology, the highly successful hot big bang (HBB) model, suffers from various fine-tuning problems (Dicke and Peebles, 1979; Linde, 1984). They claim that the spacetimes on which the HBB model is based, the Friedman-Robertson-Walker (FRW) spacetimes, require seemingly “special” initial conditions, such that when they are evolved forward in time by the dynamical law of the general theory of relativity (GTR) they yield presently observed cosmological conditions. For example, the flatness problem depends on the existence of special initial conditions in the HBB model which are required to explain the observationally-inferred spatial flatness of the universe. Due to their extreme precision or intuitive “unlikeliness,” these initial conditions are thought to be unduly special, such that many cosmologists have felt that the initial conditions themselves are in need of explanation and, moreover, present a significant conceptual problem for the HBB model.

Although physical fine-tuning could be interpreted in a variety of ways, cosmologists typically understand it to mean that observationally-required initial conditions are in some sense unlikely (Smeenk, 2013; McCoy, 2015). In order to substantiate this interpretation, one must show that initial conditions in the HBB model which reproduce present conditions are in fact unlikely. This task presupposes that there is a justifiable way of assessing the likelihoods of cosmological models (Gibbons et al., 1987; Hawking and Page, 1988). Many arguments found in the cosmological literature, however, rely on ad hoc, unjustified likelihood measures. Gibbons et al. (1987) propose a “natural” measure (hence the GHS measure) on the set of FRW spacetimes (with matter contents represented by a scalar field) as a natural and justified way of evaluating likelihoods. The GHS measure is simply the canonical Liouville measure associated with the phase space of FRW spacetimes when GTR is put into a Hamiltonian formulation and in a precise sense “comes for free” with the phase space.

While I would maintain that the GHS measure cannot be successfully used to make arguments about fine-tuning in cosmology quite generally, I argue here only for its inapplicability to the flatness problem. Some

*Eidyn Research Centre, University of Edinburgh, Edinburgh, UK. email: casey.mccoy@ed.ac.uk

authors (Gibbons and Turok, 2008; Carroll and Tam, 2010) have attempted to make probabilistic arguments, in analogy to familiar probabilistic arguments in statistical mechanics, by making the GHS measure into a probability measure. However, as the total measure of the FRW phase space is infinite, there is no canonical choice of probability measure with which to make probabilistic arguments, a point that has been recognized already by some (Hawking and Page, 1988; Schiffrin and Wald, 2012). Accordingly, any justification of a particular probability measure is completely independent of the justification of the GHS measure—in short, these probability measures are not in any substantive sense the GHS measure. On the other hand, one might try to use the GHS measure by itself to make typicality arguments in analogy to typicality arguments in statistical mechanics (Goldstein, 2012). Carroll in particular advocates this approach and, interestingly, claims that the GHS measure alone tells us that almost all spacetimes are spatially flat (Carroll and Tam, 2010; Remmen and Carroll, 2013; Carroll, forthcoming)—that there is in fact no flatness problem (Hawking and Page (1988, 803-4) and Coule (1995, 468) suggest the same). Carroll’s claim, however, rests on a subtle mistake in interpreting typicality. I claim, on the contrary, that the GHS measure cannot tell us anything about likelihood without substantive additional assumptions such as those made in statistical mechanics, e.g. a partition of phase space into “macroproperties” or similar. These necessary assumptions, however, are doubtfully justifiable in the cosmological context. Thus I ultimately conclude that the GHS measure cannot be used to clarify the nature of fine-tuning in cosmology.

1 The Gibbons-Hawking-Stewart Measure

An adequate view of what the GHS measure is and can do relies on understanding the details of how it is introduced. For this reason I develop here the measure with considerably more care than other accounts in the literature, which tend to jump straight to a Lagrangian or Hamiltonian formulation of GTR without elucidating the geometrical origin of their variable choices and the relations between physical parameters.

My starting point is the initial value formulation of GTR, in which the “position” initial data of spacetime are represented by the spatial metric h_{ab} on a spacelike Cauchy surface Σ and the “momentum” initial data by the extrinsic curvature π_{ab} (Wald, 1984; Malament, 2012). FRW spacetimes are spacetimes with homogeneous and isotropic spacelike hypersurfaces, so one can foliate the spacetimes by a one-parameter family of these spacelike hypersurfaces Σ_t that are orthogonal to a smooth, future-directed, twist-free, unit timelike field ξ^a on M , where I define $\xi^a = \nabla^a t$. For FRW spacetimes the extrinsic curvature of an initial data surface Σ_t is Hh_{ab} , where H is the so-called Hubble parameter. Thus the initial data for an FRW spacetime are completely represented by two objects: (1) the spatial metric h_{ab} and (2) the Hubble parameter H associated with a spatial hypersurface Σ .

The space of initial data is therefore the product of the set of homogeneous and isotropic Riemannian manifolds Σ (with metric h_{ab}) and the set of (real-valued) Hubble parameters H . Homogeneous and isotropic Riemannian manifolds have constant curvature κ . Complete, connected Riemannian manifolds of constant sectional curvature are called space forms. It is a theorem that every simply-connected three-dimensional space form is isometric to the sphere $S^3(\sqrt{1/\kappa})$ if $\kappa > 0$, \mathbf{R}^3 if $\kappa = 0$, or the hyperbolic space $H^3(\sqrt{1/\kappa})$ if $\kappa < 0$ (Wolf, 2010). The standard metrics on each of these manifolds is understood to be the metric induced on them by embedding them in \mathbf{R}^4 . Every Σ is therefore isometric to one of these three classes of space forms. Spaceforms of each of the three kinds are moreover homothetic, i.e. they are isometric up to the square of a scale factor a (McCabe, 2004). Accordingly one has the means to represent curvature κ as a function of the scale factor; in particular, for any Σ , $a^2\kappa$ is some constant k . Hence one can set any spatial metric $h_{ab} = a^2\gamma_{ab}$, where γ_{ab} is the standard metric on the appropriate space form. This is useful in the initial value formulation of FRW spacetimes because all time dependence of h_{ab} is thereby located solely in

the scale factor rather than in the radius of curvature of the space form.

The Einstein equation reduces to two constraint equations and two evolution equations in the initial value formulation (Geroch, 1972):

$$\mathcal{R} - (\pi_a^a)^2 + \pi_{ab}\pi^{ab} = -16\pi T_{ab}\xi^a\xi^b; \quad (1)$$

$$D_c\pi_a^c - D_a\pi_c^c = 8\pi T_{mr}h_a^mh_b^r\xi^r; \quad (2)$$

$$\mathfrak{L}_\xi(\pi_{ab}) = 2\pi_a^c\pi_{cb} - \pi_c^c\pi_{ab} + \mathcal{R}_{ab} - 8\pi h_a^mh_b^n(T_{mn} - \frac{1}{2}Th_{mn}); \quad (3)$$

$$\mathfrak{L}_\xi(h_{ab}) = 2\pi_{ab}, \quad (4)$$

where \mathcal{R} is the Ricci scalar of Σ , \mathcal{R}_{ab} is the Ricci tensor of Σ , and D_a is the derivative operator on Σ . For FRW spacetimes, these equations simplify to the following three (the second equation from above is trivial since π_{ab} does not vary across Σ):

$$\mathcal{R} - 6H^2 = -16\pi\rho; \quad (5)$$

$$\dot{H}h_{ab} = \left(-H^2 - \frac{4\pi}{3}(\rho + 3p)\right)h_{ab}; \quad (6)$$

$$\dot{h}_{ab} = 2Hh_{ab}, \quad (7)$$

where ρ is the energy density and p the pressure of the matter. The first two equations are known as the Friedman equations. Since $h_{ab} = a^2\gamma_{ab}$, $\dot{h}_{ab} = 2a\dot{a}\gamma_{ab}$, and $2Hh_{ab} = 2Ha^2\gamma_{ab}$, it follows from the third equation above that

$$H = \frac{\dot{a}}{a}, \quad (8)$$

which is the usual definition of the Hubble parameter H . To simplify matters somewhat and to make contact with the literature, I shall henceforth take the matter contents of spacetime to be a scalar field ϕ in a potential V which evolves according to the coupled Einstein-Klein Gordon equation.¹ Then one has the following equations of motion (Hawking and Page, 1988, 790):

$$\mathcal{R} - 6H^2 = -16\pi\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right) \quad (9)$$

$$\dot{H} = -H^2 - \frac{8\pi}{3}\left(\frac{1}{2}\dot{\phi}^2 - V(\phi)\right) \quad (10)$$

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0, \quad (11)$$

where V' is the derivative of the potential with respect to ϕ .² (The third equation can be derived from the previous two, and so is in fact redundant.)

For FRW spacetimes the spatial Ricci scalar is $\mathcal{R} = -6\kappa$. As noted before, one can cast κ in terms of the scale factor and a constant k : $\kappa = k/a^2$. By using the scale factor a to replace κ , one has introduced a constant k which has no physical significance beyond identifying whether the space form is flat, positively-curved, or negatively-curved. One therefore usually takes equivalence classes of curves according to these three cases and chooses $k = +1, 0$, and -1 as representatives. Then one may write $\mathcal{R} = -6k/a^2$, so that one finally has Friedman's equation in its usual form (for a scalar field in a potential):

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right) - \frac{k}{a^2}. \quad (12)$$

¹The scalar field is meant to be the inflaton, the field that drives inflation in the early universe.

²If our interest were solely in assessing the HBB model's fine-tuning, one could do the following analysis for perfect fluid matter contents. The results would be qualitatively similar however, as shown by Carroll and Tam (2010, §4.2).

The foregoing indicates that our FRW initial data h_{ab} and π_{ab} are equivalently representable in the space $\{a, \dot{a}, \phi, \dot{\phi}, k\}$. This space is not the space of initial data, however, since the previous equation is a constraint that must be satisfied by initial data. One must also keep in mind that k is an index for three separate copies of the space $\{a, \dot{a}, \phi, \dot{\phi}\}$. There is no continuous path between the three spaces.

Have identified the relevant spaces for representing FRW space forms, I next put the theory into a Hamiltonian formulation (Wald, 1984, Appendix E) in order to obtain a symplectic structure and, hence, the canonical measure. I begin with the Lagrangian for our theory of FRW spacetimes with a scalar field as the matter contents, where I have re-introduced the lapse function N as a Lagrange multiplier:

$$\mathcal{L} = \sqrt{-g} \left(\frac{R}{16\pi} + \frac{1}{2N^2} \dot{\phi}^2 - V(\phi) \right). \quad (13)$$

In terms of the variables I have chosen, this is

$$\mathcal{L} = -\frac{1}{8\pi} \left(\frac{3}{N} a \dot{a}^2 - 3Na^3 \frac{k}{a^2} \right) + \frac{1}{2N} a^3 \dot{\phi}^2 - Na^3 V(\phi), \quad (14)$$

in agreement with (Hawking and Page, 1988; Gibbons and Turok, 2008; Carroll and Tam, 2010). The momenta of a and ϕ are

$$p_a \equiv \frac{\partial \mathcal{L}}{\partial \dot{a}} = \frac{-3a\dot{a}}{4\pi N}; \quad p_\phi \equiv \frac{\partial \mathcal{L}}{\partial \dot{\phi}} = \frac{a^3 \dot{\phi}}{N}. \quad (15)$$

The Hamiltonian on this phase space is

$$\mathcal{H} = p_a \dot{a} + p_\phi \dot{\phi} - \mathcal{L} = N \left(-\frac{2\pi p_a^2}{3a} + \frac{p_\phi^2}{2a^3} + a^3 V(\phi) - a^3 \frac{3}{8\pi} \frac{k}{a^2} \right), \quad (16)$$

from which one recovers (after setting $N = 1$) our constraint (the Friedman equation) as the Hamiltonian constraint C :

$$C \equiv -\frac{2\pi p_a^2}{3a} + \frac{p_\phi^2}{2a^3} + a^3 V(\phi) - a^3 \frac{3}{8\pi} \frac{k}{a^2} = 0. \quad (17)$$

The phase space γ of our system is thus the four-dimensional space $\{a, p_a, \phi, p_\phi\}$ equipped with the canonical symplectic form

$$\omega_{p_a, a, p_\phi, \phi} = dp_a \wedge da + dp_\phi \wedge d\phi. \quad (18)$$

The dynamically accessible phase space points are constrained to be on the three-dimensional hypersurface C . Thus it would be inappropriate to use ω for constructing a canonical volume measure on phase space. One can, however, pull the symplectic form back onto the constraint surface by first solving the constraint for p_ϕ :³

$$p_\phi = a^3 \left(\frac{4\pi p_a^2}{3} + \frac{3}{4\pi} \frac{k}{a^2} - 2V(\phi) \right)^{1/2}. \quad (19)$$

Following Carroll and Tam, I also switch coordinates from p_a to H , so that

$$p_\phi = a^3 \left(\frac{3}{4\pi} (H^2 + k/a^2) - 2V(\phi) \right)^{1/2} \quad (20)$$

³The scalar field can have positive or negative momentum, so strictly speaking there should be a \pm in the following equation. The reader is welcome to annotate the equations that follow.

and

$$dp_a = -\frac{3}{4\pi}(2aHda + a^2dH). \quad (21)$$

The differential of p_ϕ is then

$$dp_\phi = \frac{(3/4\pi)a^3HdH - a^3V'd\phi + 6a^2((3H^2 + 2k/a^2)/8\pi - V)da}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}. \quad (22)$$

Substituting these into ω then gives the pullback of the symplectic form onto C . The result is the following (pre-symplectic) differential form:

$$\omega_{a,H,\phi} = \Theta_{Ha}(dH \wedge da) + \Theta_{H\phi}(dH \wedge d\phi) + \Theta_{a\phi}(da \wedge d\phi), \quad (23)$$

where

$$\Theta_{Ha} = -\frac{3}{4\pi}a^2; \quad (24)$$

$$\Theta_{H\phi} = \frac{(3/4\pi)a^3H}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}; \quad (25)$$

$$\Theta_{a\phi} = \frac{6a^2((3H^2 + 2k/a^2)/8\pi - V)}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}. \quad (26)$$

This form is not symplectic (it is degenerate), so one cannot construct a natural volume measure on C . Ideally, the “real” phase space of our system would be given by “solving the dynamics,” and then taking equivalence classes of phase points that are part of the same trajectory. In this way one would obtain the space of motions, onto which one could then pull back the degenerate form to obtain a new symplectic form (of degree two less than ω) and construct a canonical measure. This is quite complicated in general due to the differential equation that must be solved. The usual approach to take instead is to set H to some value H_* in the differential form and define their measure accordingly, i.e. set

$$d\Omega = \omega_{a,H,\phi}|_{H=H_*} = \Theta_{a\phi}|_{H=H_*}dad\phi. \quad (27)$$

One may do this because surfaces of constant Hubble parameter in phase space are transverse to temporal evolution, and the measure is preserved under translation of these surfaces along the Hamiltonian flow. Finally, one may naturally define the GHS measure μ_{GHS} on Lebesgue measurable sets U by

$$U \mapsto \int_U d\Omega = -6 \int_U a^2 \frac{(3H_*^2 + 2k/a^2)/8\pi - V}{((3/4\pi)(H_*^2 + k/a^2) - 2V)^{1/2}} dad\phi. \quad (28)$$

This expression of the GHS measure is equivalent to those derived in (Carroll and Tam, 2010; Schiffrin and Wald, 2012).⁴

⁴There are some complications with the $k = 1$ case. See (Schiffrin and Wald, 2012, 8) for the details. I have however chosen not to set $8\pi G = 1$, but rather maintained consistency with the rest of this dissertation’s use of “geometrical units” by only setting $G = 1$. Gibbons et al. (1987) use a simplifying, but less transparent coordinate choice. They also choose to investigate only the special case where $V = m^2\phi^2/2$. It can be shown with some work that their expression is equivalent to this one as well with this potential.

2 The Flatness Problem

The GHS measure clearly diverges for large scale factors, a point originally recognized by Gibbons et al. (1987, 745); it also converges to 0 for small scale factors. Due to the divergence, one may readily say that, given any choice of Hubble parameter H_* , almost all spacetimes will have a “large” scale factor. More precisely, pick any scale factor a_* ; the set of spacetimes with $a < a_*$ is a negligible set: the total measure of this set is finite whereas the total measure of its complement is infinite. What is the significance of this fact about the GHS measure, specifically for the flatness problem?

Hawking and Page (1988, 803-4) suggest the following:

“Thus for arbitrarily large expansions (and long times), and for arbitrarily low values of the energy density, the canonical measure implies that almost all solutions of the Friedmann-Robertson-Walker scalar equations have negligible spatial curvature and hence behave as $k = 0$ models. In this way a uniform probability distribution in the canonical measure would explain the flatness problem of cosmology...”

By “arbitrarily large expansions” (and “arbitrarily low values of energy density”), they appear to mean the following. Pick any arbitrary a_* (and any arbitrary ϕ_*).⁵ According to the GHS measure almost all spacetimes have $a > a_*$ (and $\phi > \phi_*$), or, equivalently, the spacetimes with $a < a_*$ (and $\phi < \phi_*$) compose a negligible set. Furthermore, since this holds for any choice of a_* , one may infer that almost all spacetimes are arbitrarily close to having $\kappa = 0$ (since $\kappa = k/a^2$) in exactly the same sense. It is perhaps somewhat misleading to say that curved FRW spacetimes with large scale factors “behave as $k = 0$ models;” the curvature does not change in such models. It is, however, surely false to say that a “uniform probability distribution” with respect to the GHS measure would explain the flatness problem of cosmology. There is in fact no such uniform probability distribution, since the GHS measure is not finite. Moreover, there is also no canonical probability distribution ρ at all which would make $U \mapsto \int_U \rho d\Omega_{GHS}$ into a probability measure—one has to make a choice in order to obtain a probability measure in the case of infinite total measure, a choice which appears completely arbitrary in this context.

Carroll and Tam (2010, 14) invite us to consider the question in more “physically transparent” terms by looking at the curvature κ , which I previously exchanged in favor of the scale factor a when deriving the GHS measure. One can recast the scale factor a as the curvature κ using the relation from before, namely $\kappa = k/a^2$. (Note especially that this switch maps the entire set of scale factors for the $k = 0$ case to the single point $\kappa = 0$.) One then defines the GHS measure (at least for curved FRW spacetimes) by the map

$$U \mapsto \int_U d\Omega = -6 \int_U \frac{1}{|\kappa|^{5/2}} \frac{(3H_*^2 + 2\kappa)/8\pi - V}{((3/4\pi)(H_*^2 + \kappa) - 2V)^{1/2}} d\kappa d\phi. \quad (29)$$

It is clear that the measure diverges for small values of curvature, i.e. curvatures close to flat, due to the curvature term in the denominator. This is pointed out by Carroll and Tam (2010, 15). They suggest the following interpretation of this fact:

“Considering first the measure on purely Robertson-Walker cosmologies (without perturbations) as a function of spatial curvature, there is a divergence at zero curvature. In other words, curved [FRW] cosmologies are a set of measure zero—the flatness problem, as conventionally understood, does not exist.”

⁵Gibbons and Turok (2008, 6) point out that ϕ is always bounded given H_* , so it is not really necessary to pick an arbitrary ϕ_* .

As stated these claims are highly suspect.

Firstly, Carroll and Tam assert that all values of their curvature coordinate Ω_k (essentially equivalent to κ) can be integrated over. While this is perhaps true, portraying the phase space in terms of curvature is misleading. For curved FRW spacetimes, it is true that the measure diverges for small values of curvature κ , as I indicate above and as Hawking and Page suggest in the passage from their paper quoted above. The recast measure, however, is infinite *at* zero curvature because the entire set of $k = 0$ scale factors is mapped to $\kappa = 0$. The GHS measure diverges for large scale factors in the case of flat FRW spacetimes just as it does for curved FRW spacetime. Thus it is misleading to describe a “divergence at zero curvature;” there is nothing special going on in flat FRW spacetimes (at least in this respect).⁶

Secondly (and relatedly), curved FRW spacetimes are clearly not a set of measure zero—at least according to the GHS measure. The initial data of FRW spacetimes is representable in the space $\{a, \dot{a}, \phi, \dot{\phi}, k\}$. The curvature constant k serves as an index for *three different phase spaces*, each of which has an infinite total measure—even after taking into account constraints and choosing a hypersurface in the constraint surface according to GHS’s procedure. The unboundedness of the total phase space measure for each kind of FRW spacetime is due, again, to the unbounded range of the scale factor Schiffrin and Wald (2012, 11).⁷ This is quite plain when one expresses the GHS measure in terms of the scale factor. Transforming to the curvature coordinate κ should not change the fact that the total measure of each phase space is infinite. So, while it is true that the GHS measure attributes infinite measure to flat FRW spacetimes (as Carroll and Tam appear to recognize), it also does so both to positively curved FRW spacetimes and to negatively curved spacetimes. Therefore it is false that the curved FRW cosmologies are a set of measure zero according to the GHS measure; hence one cannot conclude on this basis that the flatness problem does not exist.

One might try to rescue Carroll and Tam’s claim about the flatness problem by interpreting flatness more broadly, namely by including “nearly flat” curved spacetimes. This requires specifying what the set of “nearly flat” curved spacetimes is to be, e.g. a specification of the set of spacetimes with curvature less than some κ_* (at some time corresponding to Hubble parameter H_*). Almost all spacetimes will have a “small” curvature κ in comparison to this curvature κ_* . In other words, the set of spacetimes with $\kappa > \kappa_*$ is a negligible set. Since our universe’s spatial curvature is thought to be “nearly flat,” i.e. it should be less than κ_* (whatever it is), it follows from this argument that our universe is actually typical, *contra* what is assumed in the flatness problem. Unfortunately this argument does not follow from the GHS measure alone, since one had to make an independent choice in choosing κ_* , a choice that is not natural in any clear sense whatever. Furthermore, it is doubtful that there is any reasonable argument to justify a choice of κ_* —an explication of “close to flat” in the context of FRW models; it appears to be a completely arbitrary choice.

Here is a slightly different tack into the same stiff headwind. Suppose κ_* is the (non-zero) spatial curvature of our universe at the present time. The GHS measure can be used to infer that almost all spacetimes with the same Hubble parameter will have flatter spatial curvatures. In such circumstances, one might be inclined to wonder “Why is my universe’s spatial curvature so large? It seems like it ought to be much smaller if my universe is typical!” On this line of thought, it seems like one actually has a curvature problem rather than a flatness problem. Of course one would say this for any κ_* whatsoever, regardless of its magnitude,

⁶Carroll and Tam appear to equivocate several times between there being a divergence *at* $\kappa = 0$ and the measure diverging *as* $\kappa \rightarrow 0$: “The integral diverges near [$\kappa = 0$], which is certainly a physically allowed region of parameter space” (Carroll and Tam, 2010, 17); “The measure diverges on flat universes” (Carroll and Tam, 2010, 28).

⁷Besides in (Schiffirin and Wald, 2012), this fact is correctly pointed out in (Gibbons et al., 1987; Hawking and Page, 1988). While Carroll and Tam (2010, 20-1) observe that “this divergence was noted in the original GHS paper, where it was attributed to ‘universes with very large scale factors’ due to a different choice of variables,” they object to this as an interpretation: “This is not the most physically transparent characterization, as any open universe will eventually have a large scale factor.” For this reason they exchange the scale factor for curvature; it is not clear, however, how this characterization is more physically transparent since it amounts to the same thing.

so it is not clear how one would ever be in the position to be satisfied with one's curvature in an FRW universe—at least insofar as one expects things in our universe to be typical (in accord with Copernican principle-style reasoning). No matter. The measure suggests this question. What is the answer?

The answer is that the curvature depends on the actual dynamical history of the universe, and so it has no explanation within the context of the HBB model (apart from one depending on an initial condition). That answer may be unsatisfying, but the question is a bad one anyway, driven by misleading intuitions. There is no such thing as a typical FRW spacetime, and the GHS measure is not going to explain why the universe's curvature is what it is. This kind of thinking is clearly motivated by supposing that the GHS measure can be used as a likelihood measure, as Carroll and Tam clearly do:

“When we consider questions of fine-tuning, however, we are comparing the real world to what we think a randomly-chosen history of the universe would be like” (Carroll and Tam, 2010, 11).

Some popular, specious conceptions (in physics and beyond) of statistical mechanics encourage this line of thought. Putatively successful typicality arguments in statistical mechanics (Goldstein, 2012) depend, however, not only on having a phase space measure, but also on both the dynamics of the system and on a specification of macroproperties or macrostates (defined as regions of phase space) (Frigg, 2009; Frigg and Werndl, 2012). Accordingly, any claim of fine-tuning in FRW spacetimes on the sole basis of the GHS measure (which does at least incorporate the FRW dynamics) is bound to miss the mark without additional assumptions (such as a well-motivated standard of flatness).

Gibbons and Turok (2008) take a different approach from Carroll and Tam. They correctly observe that universes with large scale factors are universes with small spatial curvatures. They then claim that the scale factor is neither “geometrically meaningful” nor “physically observable” and therefore propose to identify all the “indistinguishable” nearly flat spacetimes on the surface identified by H_* .⁸ They do so by effectively choosing a “cutoff” curvature κ_* and throwing out all the spacetimes with curvatures smaller than it. The advantage to doing this is that the total measure of FRW spacetimes with curvatures larger than κ_* is finite, so that one can then define a probability measure in a natural way.

The disadvantage is that this makes no sense. Carroll and Tam (2010, 20) comment, “to us, this seems to be throwing away almost all the solutions, and keeping a set of measure zero. It is true that universes with almost identical values of the curvature parameter will be physically indistinguishable, but that doesn't affect the fact that almost all universes have this property.” Indeed, doing what Gibbons and Turok do is throwing away almost all the solutions (although the remaining set has finite measure, not measure zero as Carroll and Tam claim). They are also right to point out that if nearly flat universes are physically indistinguishable, so are “nearly- κ ” universes for almost any κ . Gibbons and Turok do not throw out these universes however (else they would not have been left with any universes at all). Their justification for an additional assumption therefore fails.

Ironically, Carroll and Tam make essentially the same error as Gibbons and Turok, by identifying the flat and nearly flat spacetimes. Instead of throwing out all the flat and nearly flat spacetimes like the latter pair, however, the former pair throws out the complement of the flat and nearly flat spacetimes by assigning them zero measure. They then triumphantly conclude that all FRW spacetimes are essentially flat! Carroll and Tam propose to tame the remaining divergence in the GHS measure by regularizing the integral, in effect making the measure finite. The problem with doing this is that, since the GHS measure is not finite,

⁸It is not clear what they mean by “geometrically meaningful.” The scale factor is clearly geometric in the relevant sense, since it relates spaceforms of the same kind by scalings. It is moreover physically meaningful because space is expanding (or contracting) in FRW spacetimes. The precise value of a does not matter, as it can be re-scaled, but that does not undermine its meaningfulness. It is also unclear how the fact that a is physically unobservable should matter, since most features of spacetime are not observable, e.g. the metric g , the spatial curvature κ , etc. The physically relevant content of these, including the scale factor, can be inferred from observations and appropriate assumptions.

regularizing the measure makes it no longer the GHS measure, in which case any justification the measure had by its “naturalness” is lost since a choice was made.⁹ In short, one may as well have just assumed the probability distribution they end up with from the very beginning. Their stated justification for this move is pragmatic: “This non-normalizability is problematic if we would like to interpret the measure as determining the relative fraction of universes with different physical properties” (Carroll and Tam, 2010, 17). However this is obviously an inadequate justification for the propriety of their measure.

References

- Albrecht, Andreas, and Paul Steinhardt. “Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking.” *Physical Review Letters* 48: (1982) 1220–1223.
- Belinsky, Vladimir, Leonid Grishchuk, Isaak Khalatnikov, and Yakov Zeldovich. “Inflationary Stages in Cosmological Models with a Scalar Field.” *Physics Letters B* 155: (1985) 232–236.
- Carroll, Sean. “In What Sense Is the Early Universe Fine-Tuned?” In *Time’s Arrows and the Probability Structure of the World*, edited by Barry Loewer, Brad Weslake, and Eric Winsberg, Cambridge, MA: Harvard University Press, forthcoming.
- Carroll, Sean, and Heywood Tam. “Unitary Evolution and Cosmological Fine-Tuning.” ArXiv Eprint, 2010. <http://arxiv.org/abs/1007.1417>.
- Coule, David. “Canonical measure and the flatness of a FRW universe.” *Classical and Quantum Gravity* 12: (1995) 455–469.
- Dicke, Robert, and Jim Peebles. “The Big Bang Cosmology—Enigmas and Nostrums.” In *General Relativity: An Einstein Centenary Survey*, edited by Stephen Hawking, and Werner Israel, Cambridge: Cambridge University Press, 1979, chapter 9, 504–517.
- Frigg, Roman. “Typicality and the Approach to Equilibrium in Boltzmannian Statistical Mechanics.” *Philosophy of Science* 76: (2009) 997–1008.
- Frigg, Roman, and Charlotte Werndl. “Demystifying Typicality.” *Philosophy of Science* 79: (2012) 917–929.
- Geroch, Robert. “General Relativity.”, 1972. Unpublished lecture notes.
- Gibbons, Gary, Stephen Hawking, and John Stewart. “A natural measure on the Set of all Universes.” *Nuclear Physics B* 281: (1987) 736–751.
- Gibbons, Gary, and Neil Turok. “Measure problem in cosmology.” *Physical Review D* 77: (2008) 1–12.
- Goldstein, Sheldon. “Typicality and Notions of Probability in Physics.” In *Probability in Physics*, edited by Yemima Ben-Menahem, and Meir Hemmo, Berlin: Springer Verlag, 2012, chapter 4, 59–71.
- Guth, Alan. “Inflationary universe: A possible solution to the horizon and flatness problems.” *Physical Review D* 23, 2: (1981) 347–356.

⁹Carroll more recently has conceded the artificiality of regularizing: “Earlier attempts to regularize the measure, for example by considering an ϵ -neighborhood around the zero-curvature Hamiltonian constraint surface (Carroll and Tam, 2010) or by identifying universes with similar curvatures (Gibbons and Turok, 2008) have not proven satisfactory” (Remmen and Carroll, 2013, 7). He remains convinced, however, that almost all FRW spacetimes are “nearly flat:” “we should throw all of the others away and deal with flat universes.” (Carroll, forthcoming, 19), developing a measure on just these spacetimes in a later paper (Remmen and Carroll, 2014).

- Hawking, Stephen, and Don Page. "How Probable is Inflation?" *Nuclear Physics B* 298: (1988) 789–809.
- Linde, Andrei. "A New Inflationary Universe Scenario: A Possible Solution of the Horizon, Flatness, Homogeneity, Isotropy, and Primordial Monopole Problems." *Physics Letters B* 108: (1982) 389–393.
- . "The inflationary universe." *Reports on Progress in Physics* 47: (1984) 925–986.
- Malament, David. *Topics in the Foundations of General Relativity and Newtonian Gravity Theory*. Chicago: University of Chicago Press, 2012.
- McCabe, Gordon. "The structure and interpretation of cosmology: Part I—general relativistic cosmology." *Studies in History and Philosophy of Modern Physics* 35: (2004) 549–595.
- McCoy, Casey. "Does inflation solve the hot big bang model's fine-tuning problems?" *Studies in History and Philosophy of Modern Physics* 51: (2015) 23–36.
- Remmen, Grant, and Sean Carroll. "Attractor solutions in scalar-field cosmology." *Physical Review D* 88: (2013) 1–14.
- . "How many e -folds should we expect from high-scale inflation?" *Physical Review D* 90: (2014) 1–14.
- Schiffrin, Joshua, and Robert Wald. "Measure and probability in cosmology." *Physical Review D* 86: (2012) 1–20.
- Smeenk, Chris. "Philosophy of Cosmology." In *The Oxford Handbook of Philosophy of Physics*, edited by Robert Batterman, Oxford: Oxford University Press, 2013, chapter 17, 607–652.
- Wald, Robert. *General Relativity*. Chicago: University of Chicago Press, 1984.
- Wolf, Joseph. *Spaces of Constant Curvature*. Providence, RI: AMS Chelsea Publishing, 2010, 6th edition.

Invariance, Interpretation, and Motivation

Thomas Møller-Nielsen

July 2016

[Forthcoming in *Philosophy of Science (2016 Proceedings)*.]

Abstract

In this paper I assess the ‘Invariance Principle’, which states that only quantities that are invariant under the symmetries of our theories are physically real. I argue, contrary to current orthodoxy, that the variance of a quantity under a theory’s symmetries is not a sufficient basis for interpreting that theory as being uncommitted to the reality of that quantity. Rather, I argue, the variance of a quantity under symmetries only ever serves as a motivation to refrain from any commitment to the quantity in question. In the process of this discussion, I address the related but importantly distinct issue of when symmetries can be said to prompt a mathematical reformulation of the relevant theory.

1 Introduction

Take the *Invariance Principle* to be the principle that only quantities that are invariant under the symmetries of our theories are physically real.¹ It is a doctrine with a distinguished pedigree: acclaimed theorists as diverse as the physicist Paul Dirac, the mathematician Hermann Weyl, and the philosopher Robert Nozick were all apparent signatories during their respective lifetimes.² *Prima facie*, however, it is something of a mystery as to how and why the principle is supposed to work. Nevertheless, there appear to be at least some uncontroversial cases where it—or something very close to it—does work.

One such example can be found in Newtonian Gravitation Theory (NGT), i.e., the theory comprising Newton’s three laws, plus his inverse square gravitational law, governing the behaviour of point particles in Newtonian spacetime. As is well known, this theory is *Galilean invariant*. This implies, among other things, that if one takes any solution to NGT and “boosts” it—that is, uniformly alters the absolute velocity of each point particle by the same amount throughout its history—one will invariably get back a solution to NGT. Boosts, in other words, are a *symmetry* of NGT: they are transformations that invariably map solutions of the theory to solutions.

¹I draw the term from Saunders (2007). Compare also Dasgupta’s (forthcoming) “symmetry-to-reality inference”.

²See, e.g., Dirac (1930, vii), Weyl (1952, 132), and Nozick (2001, 82).

Which quantity varies under this particular symmetry? The answer is obvious: absolute velocity. Thus, according to the Invariance Principle, we should conclude that absolute velocity is not a genuine physical quantity. Conversely, which quantities are invariant under this particular symmetry? Again, the answer is obvious: relative (inter-particle) distance and velocity, temporal intervals, and absolute acceleration. Thus, according to the Invariance Principle, we should conclude that NGT's boost symmetry does not threaten these quantities' status as genuinely physical.

As it turns out, one can successfully purge Newtonian theory of the spacetime structure required to make absolute velocity a physically meaningful quantity. More specifically, one can move to *Galilean spacetime*. (Sometimes also called "Neo-Newtonian spacetime".)³ Here, the Newtonian posit of persisting points of absolute space—persisting points which, crucially, allow for the notion of absolute velocity to be physically meaningful—is done away with, but an *affine structure* is nevertheless preserved, which defines the "straight" or force-free (inertial) paths through spacetime. Absolute velocity is therefore not a physically meaningful quantity in Galilean spacetime, as it is in Newtonian spacetime. Nevertheless, all other Newtonian notions, including the notion of absolute acceleration, remain well-defined in Galilean spacetime. To the extent that one opts for Galilean over Newtonian spacetime, then, one has excised an ostensibly odious piece of theoretical structure from NGT.

Three important caveats are worth noting, however. First, and most obviously, none of this is to say that Newtonian theory set in Galilean spacetime is therefore the true and complete theory of the world. (It isn't.) Second, nor is this to say that by moving to Galilean spacetime one has thereby purged Newtonian theory of all its "variant" structure. (One hasn't. The symmetry group of Newtonian theory is actually wider than the Galilean group: it has additional symmetries.)⁴ Third, nor is this even to say that the invariant quantities one ends up with following such an application of the Invariance Principle will invariably be preserved in future theories. (For instance, there is no notion of "relative spatial distance" *simpliciter* in special relativity.) Given all of these caveats, however, one might well ask: What good is the Invariance Principle, exactly? What purpose, in particular, does it serve?

As I see it—and, I take it, as many other contemporary theorists also see it—the purpose of the Invariance Principle is essentially *comparative*. That is, it is simply supposed to lead you to a *better theory*—or a better interpretation, or characterisation, of the same theory—than the one you started with. To take the case at hand: Newtonian theory set in Galilean spacetime is a better theory than Newtonian theory set in Newtonian spacetime. It is a theory which possesses all of the theoretical virtues of its rival, but lacks any apparent ontological commitment to the unwanted variant quantity in question.

In summary, the Galilean invariance of NGT, in conjunction with the Invariance Principle, is supposed to indicate that neither absolute velocity nor

³See, e.g., Earman (1989, §2.4).

⁴See, e.g., Knox (2014). I discuss this point further in Section 4 below.

any corresponding persisting points of absolute space are genuinely real. Now to lay my cards on the table: I actually think that something *very close* to this general kind of inference—that is, from the variance of a quantity under symmetries to that quantity’s nonreality—is legitimate. The devil, however, is in the details. In particular, I don’t believe that the *mere* Galilean invariance of NGT is enough to establish absolute velocity’s nonreality. And in general, I don’t believe that the *mere* variance of a quantity under symmetries is enough to establish that quantity’s nonreality. These beliefs, as far as I can determine, put me in the minority camp in the contemporary philosophical literature on symmetries. Nevertheless, I think they are correct beliefs—and they are precisely the ones that I will attempt to argue for in the remainder of this paper.

2 Interpretational vs Motivational

In arguing for the above claims, it will prove extremely useful first to distinguish between two very different ways of thinking about symmetries.

Close cousins of the distinction that I have in mind have already been drawn in the literature. Thus, Greaves and Wallace write:

There is a widespread consensus that two states of affairs related by a symmetry transformation are really just the same state of affairs differently described. That is, if two mathematical models of a physical theory are related by a symmetry transformation, then those models represent one and the same physical state of affairs. (Greaves and Wallace 2014, 60)

They continue:

Although we agree with this consensus [...] even those who do not agree that symmetry-related states of affairs are identical at least agree that they are *empirically indistinguishable* from one another. (Greaves and Wallace 2014, 60, fn 1)

To illustrate the difference between these two ways of thinking about symmetries, consider again the example of boosts in NGT. According to the “widespread consensus” view alluded to, and endorsed by, Greaves and Wallace, boosted models of NGT are to be taken to represent the same physical state of affairs *even when the theory is putatively set in Newtonian spacetime*. In other words, according to this view, one needn’t make the move to Galilean spacetime in order not to be committed to absolute velocities; there is a way of understanding boosted models’ physical equivalence, and their associated noncommitment to the notion of absolute velocity, prior to making this move.⁵

Things are very different according to the second conception of symmetries described, and rejected, by Greaves and Wallace. According to this view, boosted models of NGT are to be regarded as physically *inequivalent*: they are not to be construed as representing the same physical state of affairs. Instead,

⁵See, e.g., Healey (2007, 114-7), for an endorsement of this view in the Newtonian context.

such models are taken to represent physically distinct scenarios, which differ in what absolute velocity they ascribe to the world's total material content. Nevertheless, such models still represent *empirically indistinguishable* states of affairs: in a Newtonian universe, no experiment could ever help an observer determine what her absolute velocity actually is. Such boosted models therefore represent physically distinct ways for the world to be, albeit ones that are indiscernible on the basis of measurement.⁶

As previously mentioned, this distinction between different ways of thinking about symmetries is close, but not identical, to the one that I want to draw. The key reason why it is not identical is because Greaves and Wallace say nothing to the effect that the person who subscribes to the second conception of symmetries—that is, who believes that symmetry-related models invariably represent empirically indistinguishable, but not necessarily physically equivalent, states of affairs—should still be *motivated to seek* an alternative theory, or an alternative interpretation or characterisation of the same theory, according to which such models do not merely represent empirically indistinguishable scenarios, but rather represent physically equivalent states of affairs.⁷ Moreover, I claim, it is precisely this notion of *motivation* which plays a central role in correctly understanding the philosophical significance of symmetries in the general case.⁸

Here, then, is what I take to be the appropriate distinction between these two different ways of thinking about symmetries:

- **Interpretational:** Symmetries allow us to *interpret* theories as being committed solely to the existence of invariant quantities, even in the absence of a metaphysically perspicuous characterisation of the reality which is alleged to underlie symmetry-related models.
- **Motivational:** Symmetries only *motivate* us to find a metaphysically perspicuous characterisation of the reality which is alleged to underlie symmetry-related models, but they do not allow us to interpret that theory as being solely committed to the existence of invariant quantities in the absence of any such characterisation.

The central claim of this paper may now be neatly summarised: the (orthodox) interpretational view is mistaken; the (unorthodox) motivational view is correct.

Drawing the distinction in the way that I have done, however, invites the rather obvious question: What, precisely, is meant by a “metaphysically perspicuous characterisation” of reality? This is the question addressed in the next section.

⁶See, e.g., Maudlin (1993, 192), for an endorsement of this view in the Newtonian context.

⁷Compare (again) Maudlin's (1993, 192) discussion in the Newtonian context.

⁸Note that I do not intend any of this as a criticism of Greaves and Wallace's paper. Indeed, as Greaves and Wallace (2014, 60, fn 1) are careful to remark, the distinction they draw is orthogonal to the central topic of their paper, namely the issue of which symmetries have “direct empirical significance” (i.e., have analogues to Galileo's ship).

3 More on Metaphysical Perspicuity

In intuitive terms, a metaphysically perspicuous characterisation of reality is one which corresponds to, or “limns”, reality’s structure in some suitably faithful way. To use another common (Platonic) metaphor, a metaphysically perspicuous characterisation of reality is one which “carves nature at its joints”. (In comparative terms: a description of reality is *more* metaphysically perspicuous than another precisely to the extent that it corresponds to, or limns, reality’s structure *more* faithfully than its rival does.)

As many readers will be aware, such a notion is frequently alluded to, and made use of, in contemporary analytic metaphysics.⁹ But metaphysical perspicuity is also, I think, a notion that is reasonably serviceable in physical (rather than “merely metaphysical”) contexts. One particularly illustrative example—albeit a slightly misleading one, for reasons that I will soon explain—drawn from physics may plausibly be found in classical electromagnetism.¹⁰ As is well known, this theory may be formulated in two different ways.¹¹ According to one such formulation, EM₁, the theory is expressed in terms of the Faraday tensor, F_{ab} , satisfying the (Maxwell) equations $\nabla_{[a}F_{bc]} = 0$ and $\nabla_a F^{ab} = J^a$, where J^a is a vector field representing the charge current density. According to the second formulation, EM₂, however, the theory is expressed in terms of the vector potential, A_a , satisfying the equation $\nabla_a \nabla^a A^b - \nabla^b \nabla_a A^a = J^b$.

These two formulations of electromagnetism are related to one another. In particular, any model $\langle M, \eta_{ab}, A_a \rangle$ of EM₂ corresponds to a unique model $\langle M, \eta_{ab}, F_{ab} \rangle$ of EM₁, via the equation $F_{ab} = \nabla_{[a}A_{b]}$. The converse, however, is not true. That is, a typical model of EM₁ does *not* typically correspond to a unique model of EM₂. More specifically, if $\langle M, \eta_{ab}, A_a \rangle$ is a model of EM₂ corresponding to a model $\langle M, \eta_{ab}, F_{ab} \rangle$ of EM₁, then so will any other model of EM₂ $\langle M, \eta_{ab}, A'_a \rangle$, where A'_a is related to A_a by a “gauge transformation” $A'_a = A_a + \nabla_a \chi$, where χ is some smooth scalar field.

It is EM₁ which, I take it, constitutes the metaphysically perspicuous characterisation of this theory. That is, it is the tensor F_{ab} which faithfully represents the fundamental ontology of the theory, namely the electromagnetic field. Not so EM₂. This second formulation may, of course, be useful for various calculational or heuristic purposes. But the key point is that the vector potential A_a *does not directly represent a genuinely real field*: rather, it is merely a mathematically convenient “shorthand” way of characterising and determining the values of the Faraday tensor, which *is* taken to represent the genuine material ontology of the theory.¹² Moreover, it is precisely by construing the vector potential in this

⁹See, e.g., O’Leary-Hawthorne and Cortens (1995, 154-7).

¹⁰Here and below, I take this theory to be set in Minkowski spacetime. Thus, the spacetime models of this theory are of the form $\langle M, \eta_{ab} \rangle$, where M is a four-dimensional differentiable manifold, and η_{ab} is the Minkowski metric.

¹¹For a recent, intriguing study of the relationship between these two different formulations of electromagnetism, see Weatherall (forthcoming). I draw heavily on his discussion over the next couple of paragraphs.

¹²Modulo, that is, certain concerns that arise as a result of the Aharonov-Bohm effect. See, e.g., Healey (2007).

way which plausibly allows us to explain and understand, in a fully transparent way, gauge-symmetry models' physical equivalence in EM_2 —namely, for the reason that they are merely notationally distinct ways of representing the same fundamental physical ontology.

As mentioned above, I think this example of metaphysical perspicuity is apt to be slightly misleading, at least when taken on its own. This is because this example might make it seem as though having a metaphysically perspicuous characterisation of the (putative) reality underlying symmetry-related models crucially relies upon one having to *mathematically reformulate* the relevant theory (or at least upon having such a mathematical reformulation already in hand), and in particular upon having to reformulate the theory so as to remove any relevant representational redundancy. However, I think this is incorrect. That is, I believe that one *can*, in fact, be in possession of a metaphysically perspicuous characterisation of the reality underlying symmetry-related models *even in the absence* of any mathematical (re-)formulation of the theory which removes the relevant representational redundancy.

Let me illustrate this point with two simple examples. First, consider the case of *shift symmetry* in NGT. This symmetry is subtly different from the case of boost symmetry, discussed above. Here, instead of uniformly altering the absolute velocity of each particle throughout its history, one enacts a global, time-independent repositioning of all matter in space. Thus, for instance, in the shifted world all of the world's material content will (*prima facie*) be located three metres to the left of where it is in the original world. The basic idea behind the “Leibniz shift” argument—the famous argument associated with this symmetry—is that the substantialist's admission of points of space as primitive objects (allegedly) has the undesirable consequence of committing her to regarding shifted worlds as physically distinct, yet nevertheless empirically indistinguishable:¹³ in intuitive terms, everything would look, feel, taste, touch and sound the same in the two (putatively distinct) shifted worlds, just as in the case of boosted worlds.

It will prove helpful to express all of this in terms of the models of the theory. Thus, take a generic model of NGT to be of the form $\mathcal{M} = \langle M, t_{ab}, h^{ab}, \sigma^a, \rho, \phi \rangle$, where M is a differentiable 4-dimensional manifold, t_{ab} is the temporal metric, h^{ab} is the spatial metric, σ^a is the timelike vector field whose integral curves represent the persisting points of absolute space, and ρ and ϕ represent the matter density and the gravitational potential field respectively.¹⁴ A shift symmetry can then be characterised as the application of the appropriate diffeomorphism (corresponding to a spatial translation) d so as to yield a new model $\mathcal{M}_{static} = \langle M, t_{ab}, h^{ab}, \sigma^a, d^*\rho, d^*\phi \rangle$. It is then alleged that \mathcal{M} and \mathcal{M}_{static} differ precisely

¹³Though see Maudlin (1993), who notes that there is an interesting (epistemological) sense in which shifted worlds in NGT are not indiscernible after all.

¹⁴Note that the canonical presentations of Newtonian spacetime (e.g., Earman 1989, §2.5) take the affine connection as ideologically primitive. I find such presentations unsatisfactory for historical rather than for philosophical reasons: in particular, it threatens to make the move to Galilean spacetime seem almost trivial, and the associated timelike vector field trivially superfluous. For more on this point, see Pooley (MS, §4.4-§4.5).

insofar as they each represent the world's matter content as being located at distinct places in absolute space. More specifically, such Leibniz-shifted scenarios are alleged to differ precisely with regard to which particular points of space are underlying various parts of the matter fields.

For a second example, consider *diffeomorphism symmetry* in general relativity (GR). Here, similarly, the existence of this symmetry is alleged to commit the substantialist to a plurality of physically distinct possibilities that are nevertheless empirically indistinguishable. In terms of the models of the theory: taking a generic model of GR to be of the form $\mathcal{M} = \langle M, g_{ab}, T_{ab} \rangle$ and applying an arbitrary diffeomorphism d to yield a new model $\mathcal{M}_{diff} = \langle M, d^*g_{ab}, d^*T_{ab} \rangle$ (where M is again a differentiable 4-dimensional manifold, g_{ab} is the metric tensor, and T_{ab} is the stress-energy tensor which, roughly speaking, represents the model's matter content), the two scenarios represented are alleged to differ with regard to which particular points of the spacetime manifold are underlying various parts of the metric and matter fields.¹⁵

It is my contention that neither the shift symmetry of NGT, nor the diffeomorphism symmetry of general relativity, by themselves motivate any mathematical reconstrual of the respective theories. This is because I believe there is a perfectly transparent, anti-haecceitist, “modestly structuralist”—but nevertheless fully substantialist—way of understanding such models' representational equivalence even in the absence of any such mathematical reformulation. On this view, spacetime points are construed as genuinely real, fundamental entities. However, they are “contextually individuated”: they are not to be understood as being anything more—or less—than “nodes” in the relational, geometrical structures in which they are embedded. Shifted models in NGT and diffeomorphically-related models in GR are thus to be understood as representing the same physical state of affairs precisely because the exact same pattern of relational, geometrical structures is represented as obtaining in each case. Moreover, this view denies that there are any primitive, singular (“haecceitistic”) facts about spacetime points which would even allow for a distinction between shifted or diffeomorphically-related scenarios to be coherently drawn.¹⁶

Whence the difference, then, between the case of gauge symmetry in electromagnetism on the one hand, and shift and diffeomorphism symmetry in NGT and GR on the other? I think the answer is straightforward. In the latter cases, the models in question are *isomorphic*: they represent worlds which differ at most with regard to which particular objects are playing which qualitative roles, i.e., they represent at most haecceitistically distinct possible worlds. Hence, adopting modest structuralism (which implies anti-haecceitism) about spacetime transparently collapses the number of possibilities represented by these models to one. In the former such case, however, the relevant models are *not* isomorphic—read “literally”, gauge-related models of EM_2 assign *qualitatively distinct* arrangements of the vector field over spacetime—hence adopting a modestly structuralist ontology does not by itself collapse the number of represented

¹⁵For further details see, e.g., Earman (1989, §9).

¹⁶For further defence of this view—which is sometimes also called *sophisticated substantialism* in the literature—see, e.g., Saunders (2003), Ladyman (2007), and Pooley (2013).

possibilities to one. In order to transparently understand such models' physical equivalence, then, a mathematical reformulation of the theory is required.

To summarise the claims made thus far: according to the motivational view of symmetries, one is invariably only motivated to regard symmetry-related models as physically equivalent; moreover, one is justified in regarding such models as physically equivalent only insofar as one is in possession of a metaphysically perspicuous characterisation of the reality which is alleged to underlie them. However, it is possible to be in possession of a metaphysically perspicuous characterisation of the reality underlying symmetry-related models even in the absence of a mathematical formulation of the theory which removes the relevant representational redundancy. Such a metaphysically perspicuous characterisation is possible just in case the symmetry-related models in question are isomorphic, or are naturally understood as representing at most haecceitistically distinct possibilities. In brief: symmetry-related, isomorphic models invariably do *not* motivate a mathematical reformulation of the relevant theory (modest structuralism invariably suffices); but symmetry-related, *non*-isomorphic models invariably *do*.¹⁷

4 In Defence of the Motivational View

Let us return once more to the case of NGT. As alluded to in Section 1, the symmetry group of this theory is quite large. For not only does it include transformations corresponding to global velocity boosts of solutions' matter content, but it also includes transformations corresponding to time-dependent translational accelerations of such content (so long as the gravitational potential field is also appropriately transformed). Thus, read "literally", the symmetries of this theory include transformations that map solutions to solutions that represent physically distinct, but nevertheless empirically indistinguishable, states of affairs in which a given material system is:

1. Force-free and stationary with respect to absolute space.
2. Force-free and moving at constant absolute velocity.
3. Absolutely accelerating under a gravitational force-field.

According to the interpretational conception of symmetries, we may legitimately take all of these symmetry-related solutions to in fact represent the same physical state of affairs—despite the fact that they are naturally understood as representing radically distinct physical situations. Things are very different, however, according to the motivational conception of symmetries. On this view, we are merely *motivated to regard* all such solutions as representing the same physical state of affairs, the motivation arising from the general Occamist principle that, other things being equal, our preferred scientific theories should not allow for solutions that represent physically distinct but nevertheless empirically indistinguishable possible worlds. According to the motivational

¹⁷See also Pooley (2013, 576-7) and Weatherall (forthcoming) for recent, related arguments to this effect.

view, then (and to repeat slightly), absent a metaphysically perspicuous characterisation of the reality underlying these symmetry-related models, we have no choice but to regard them as representing physically distinct states of affairs.

For our purposes, the crucial thing to note about all of these models is that *none of them are isomorphic*—naturally understood, they do not represent at most haecceitistically distinct possible worlds. According to the criterion laid down in the previous section, then, in order to be able to transparently understand how it could be that such models may be said to represent physically equivalent scenarios, a mathematical reformulation of the theory is required.

As it turns out, such a mathematical reformulation of the theory is possible. In brief, in this reformulation one replaces the vector field σ^a with a new kind of *dynamical* inertial connection ∇^{NC} , with models of the form $\mathcal{M}_{NC} = \langle M, t_{ab}, h^{ab}, \nabla^{NC}, \rho \rangle$. Up to isomorphism, any two symmetry-related models of NGT correspond to a unique model of Newtonian gravity geometrised in this way. Thus, it is said, by moving to this “Newton-Cartan” theory one successfully removes the undesirable “gauge-redundancy” inherent in all non-geometrised versions of Newtonian gravitation theory.¹⁸

What might the defender of the interpretational view of symmetries say in defence of her view—in this context, that the move to Newton-Cartan theory is not required in order to be able to legitimately regard all symmetry-related solutions of NGT as physically equivalent?

I anticipate two likely lines of response. First, she might attempt to establish the preferability of her view over the motivational view by noting that the defender of the motivational view is committed, at least prior to the appropriate theory’s reformulation (in the context of NGT), to the existence of in principle undetectable (symmetry-variant) matters of fact. Moreover, the defender of the interpretational view might argue, this is an unpalatable consequence, one which we would do best to avoid—and one which, she might point out, the interpretational view does in fact avoid.

I agree that the admission of such in principle undetectable facts is an undesirable consequence of the motivational view. However, I do not think that this admission is sufficiently unpalatable so as to be capable of refuting the motivational view, or even of establishing the preferability of the interpretational view over the motivational view. After all, prohibitively strong versions of verificationism aside, there is nothing obviously absurd about admitting in principle undetectable facts into one’s ontology; nor is there any obvious reason why we should always be capable of discovering a theory, or a perspicuous characterisation thereof (the case of isomorphic models excepted), which succeeds in transparently explaining such solutions’ empirical equivalence by virtue of

¹⁸For further details, see, e.g., Knox (2014). Note also the important point that moving to Newton-Cartan theory is not by itself sufficient for one to be able to transparently understand as physically equivalent all symmetry-related models of Newtonian theory set in flat spacetime. This is because—as mentioned above—such symmetry-related models will typically correspond to a single model of Newton-Cartan theory *only up to isomorphism*. Thus, in order to have a *fully* transparent understanding of how it is that symmetry-related models of Newtonian theory set in flat spacetime can correspond to a single model of Newton-Cartan theory, a modestly structuralist conception of spacetime ontology is also required.

their actual physical equivalence; nor indeed is there even any obvious way of guaranteeing that there will always be such a theory or characterisation (again, isomorphic models excepted) waiting in logical space to be discovered.

Furthermore, although it is to be admitted that the Newtonian who subscribes to the merely motivational view of symmetries might indeed be committed to the possibility of there being facts beyond her epistemic grasp, it nevertheless bears emphasising that for such a Newtonian there is a perfectly good explanation as to *why* such facts are epistemically inaccessible: they are inaccessible precisely because the world is in fact accurately described by the laws of NGT, with associated models of the form $\langle M, t_{ab}, h^{ab}, \sigma^a, \rho, \phi \rangle$, and because all any Newtonian observer ultimately has empirical access to are the relative distances and velocities between material entities. For such a Newtonian, then, the empirical phenomena underdetermine the genuine physical facts; but the theory itself is able to provide a perfectly transparent explanation of the reality behind the phenomena in terms of which the underdetermination can be straightforwardly understood.

The Newtonian who adopts the interpretational construal of symmetries, however, would appear to lose this explanatory transparency. In other words, she might know *that* she may legitimately regard all symmetry-related solutions as physically equivalent; but the reality in terms of which this physical equivalence is to be understood will (absent a reformulation of the theory) remain opaque to her; she is offered no immediate explanation as to *how* such physical equivalence is to be construed, or how it could even be said to arise.

These considerations naturally suggest a second possible line of response for the defender of the interpretational view. In particular, she might claim that she *does*, in fact, have a transparent understanding of the reality underlying NGT's symmetry-related models, and that such a transparent understanding is in fact attainable *prior* to the move to Newton-Cartan theory.¹⁹

Such a response evidently leads into deep philosophical waters very quickly. (After all, what does it mean to be in possession of a “transparent understanding” of anything?) But let me make a brief remark as to why I find this particular claim to be implausible. For note that in NGT the persisting points of absolute space are not merely “idly turning wheels” that can simply be expunged from the theory without explanatory loss: they are not “explanatorily idle” posits. This is for two main reasons. First, such points play a crucial role in the *metaphysical* explanation of what quantities like relative velocity and absolute rotation and absolute acceleration truly are: for the Newtonian, facts about particular inter-particle velocities and absolute rotations and absolute accelerations are naturally understood as being *grounded in* particular facts about (rates of change of) absolute velocities.²⁰ Second, such points provide the crucial transtemporal standard which is required in the realist's *causal* explanation of the observable effects of noninertial motion (e.g., Newton's famous “bucket experiment”): a standard without which Newton's laws simply cannot be formu-

¹⁹Dewar (2015, esp. 322)—who is a recent, explicit defender of the interpretational view—is plausibly read as making this claim.

²⁰Cf. Pooley (MS, 118).

lated (at least, absent any *other* way of construing the transtemporal structure required to underwrite the distinction between inertial and noninertial motion). In short—and to the extent that the interpretational view is not supposed to reduce to a rather uninteresting form of scientific instrumentalism—it is simply not clear what causal-explanatory, *realistic* picture of the world is being proffered by the defender of the interpretational view, at least in this particular (Newtonian) context; it is simply opaque what, according to her, *the world is really like*.

Acknowledgements

For extremely helpful comments and discussion, I would like to thank Neil Dewar, James Ladyman, Niels Martens, Tushar Menon, Oliver Pooley, James Read, Simon Saunders, Alex Skinner, Teru Thomas, David Wallace, and audiences in London and Cardiff.

References

- Dasgupta, S. (forthcoming), “Symmetry as an Epistemic Notion (Twice Over).” *British Journal for the Philosophy of Science*.
- Dewar, N. (2015), “Symmetries and the Philosophy of Language.” *Studies in the History and Philosophy of Modern Science*, Vol. 52, pp. 317-327.
- Dirac, P. A. M. (1930), *The Principles of Quantum Mechanics*. Oxford University Press. (Reference is made to 1958 (4th) edition.)
- Earman, J. (1989), *World-Enough and Space-Time*. MIT Press.
- Greaves, H. and Wallace, D. (2014), “Empirical Consequences of Symmetries.” *British Journal for the Philosophy of Science*, Vol. 65, No. 1, pp. 59-89.
- Healey, R. (2007), *Gauging What’s Real*. Oxford University Press.
- Knox, E. (2014), “Newtonian Spacetime Structure In Light of the Equivalence Principle.” *British Journal for the Philosophy of Science*, Vol. 65, No. 4, pp. 863-880.
- Ladyman, J. (2007), “Scientific Structuralism: On the Identity and Diversity of Objects in a Structure.” *Aristotelian Society Supplementary Volume*, Vol. 81, No. 1, pp. 23-43.
- Maudlin, T. (1993), “Buckets of Water and Waves of Space: Why Spacetime Is Probably a Substance.” *Philosophy of Science*, Vol. 68, No. 2, pp. 183-203.
- Nozick, R. (2001), *Invariances: The Structure of the Objective World*. Harvard University Press.
- O’Leary-Hawthorne, J. and Cortens, A. (1995), “Towards Ontological Nihilism.” *Philosophical Studies*, Vol. 79, No. 2, pp. 143-165.
- Pooley, O. (2013), “Substantialist and Relationalist Approaches to Spacetime.” In R. Batterman (ed.), *Oxford Handbook of Philosophy of Physics*. Oxford University Press.
- Pooley, O. (MS), *The Reality of Spacetime*. Book manuscript.

Saunders, S. (2003), "Physics and Leibniz's Principles." In K. Brading & E. Castellani (eds.), *Symmetries in Physics: Philosophical Reflections*. Cambridge University Press.

Saunders, S. (2007), "Mirroring as an A Priori Symmetry." *Philosophy of Science*, Vol. 74, No. 4, pp. 452-480.

Weatherall, J. (forthcoming). "Understanding Gauge." *Philosophy of Science*.

Weyl, H. (1952), *Symmetry*. Princeton University Press.

Black Holes, Information Loss and the Measurement Problem

Elias Okon

*Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México,
Mexico City, Mexico.*

Daniel Sudarsky

*Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico
City, Mexico.*

The *information loss paradox* is often presented as an unavoidable consequence of well-established physics. However, in order for a genuine paradox to ensue, non-trivial assumptions about, e.g., quantum effects on spacetime, are necessary. In this work we will be explicit about these additional, speculative assumptions required. We will also sketch a map of the available routes to tackle the issue, highlighting the, often overlooked, commitments demanded of each alternative. In particular, we will display the strong link between black holes, the issue of information loss and the measurement problem.

1 Introduction

The so-called *information loss paradox* is usually introduced as an unavoidable consequence of standard, well-established physics. The paradox is supposed to arise from a glaring conflict between Hawking's black hole radiation and the fact that time evolution in quantum mechanics preserves information. However, the truth is that, in order for a genuine paradox to appear, a sizable number of additional, non-standard assumptions is required. As we will see, these extra assumptions involve thesis regarding the fundamental nature of Hawking's radiation, guesses regarding quantum aspects of gravity and even considerations in the foundations of quantum theory.

In this work, we will be explicit about the additional assumptions required for a genuine conflict to arise and delineate the available options in order to tackle the issue. In particular, we will stress the connection between information loss and the measurement problem, and display the often non-trivial commitments that each of the available alternatives to solve the information loss issue demands.

2 The classical setting: black holes hide information

We start by reviewing some properties of classical black holes. Gravity, being always attractive, tends to draw matter together to form clusters. In fact, if the mass of a cluster is big enough, nothing will be able to stop the contraction until, eventually, a black hole will form. That is, the gravitational field at the surface of the body will be so strong that not even light will be able to escape and a region of spacetime from which nothing is able to emerge will form. The boundary of such a region is called the event horizon and, according to general relativity, its area never decreases.

In general, the collapse dynamics that leads to the formation of a black hole can, of course, be very complicated. However, it can be shown that all such systems eventually settle down into one of the few stationary black hole solutions, which are completely characterized by the mass, charge and angular momentum of the the Kerr-Newman spacetimes. In fact, the so-called black hole uniqueness theorems guarantee that, as long as one only considers gravitational and electromagnetic fields, then these solutions represent the complete class of stationary black holes. Moreover, the so-called no-hair theorems ensure that the set of stationary solutions does not grow, even if one considers other hypothetical fields.

The above mentioned results seem to suggest that when a cluster collapses to form a black hole, a large amount of information is lost. That is, details such as the multipole moments of the initial mass distribution, or the type of matter involved, seem to be altogether lost when the black hole settles. Note however that such apparent loss of information corresponds only to that available to observers outside of the black hole. While at early times there are Cauchy hypersurfaces¹ completely contained outside of the black hole, at later times all Cauchy hypersurfaces have parts both inside and outside it (see Figure 1). Therefore, using data located both outside and inside of the black hole, the *whole* spacetime can always be recovered. We conclude that, in the classical setting, information is not really lost. All that happens is that, when a black hole forms, a new region of no escape emerges and some of the information from the outside of the black hole moves into such new region. One could still argue that, since there are points inside of the horizon which are not in the past of future null infinity,²

¹A Cauchy hypersurface is a subset of spacetime which is intersected exactly once by every inextendible, non-spacelike curve.

²Future null infinity is the set of points which are approached asymptotically by null rays which

then it is impossible to reconstruct the whole spacetime by evolving backwards the data on it. However, future null infinity is not a Cauchy hypersurface so one should not expect to reconstruct the whole spacetime from such data.

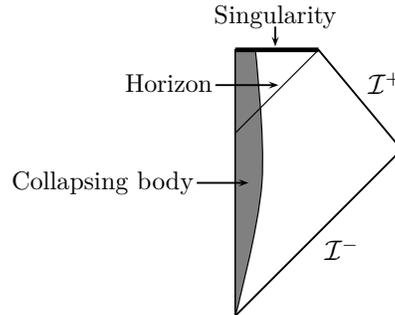


Figure 1: Penrose diagram for a collapsing spherical body. \mathcal{I}^+ and \mathcal{I}^- denote past and future null infinity.

3 QFT on a fixed curved background: black holes radiate

The most dramatic change in our understanding of black hole physics came as a result of Hawking's famous analysis. What this analysis showed was that the formation of a black hole would modify the state of any quantum field in such a way that, at late times, there would be an outgoing flux of particles carrying energy towards infinity. Moreover, Hawking showed that the flux was characterized by the surface gravity κ of the resulting asymptotic stationary state of the black hole. This discovery transformed our perception of the formal analogy, originally pointed out in Bekenstein (1972), between the laws of black hole dynamics, and the standard laws of thermodynamics (see Wald (1994) for a discussion). In particular, it led to the view that the surface gravity is in fact a measure of the black hole's temperature $T = \frac{\kappa}{2\pi}$, and that the event horizon's area A is a measure of the black hole's entropy $S = A/4$.

Hawking's result is probably the most famous of the effects that arise from the natural extension of special relativistic quantum field theory to the realm of curved spacetimes. It imposes a dramatic modification on the classical view of black holes as can escape to infinity.

absolutely black and eternal regions of spacetime. It is important to stress, though, that Hawking's calculation, being a result pertaining to quantum field theory on a *fixed* spacetime, does not encompass back-reaction effects. These are in fact notoriously difficult to deal with and a general framework for doing so is lacking. At any rate, some straightforward physical considerations, which have rather dramatic consequences, are often brought to bear in this context.

4 Back-reaction and first quantum gravity input: black holes evaporate

As can be expected, Hawking's result also suggests a dramatic modification in our expectation for the ultimate fate of a black hole. That is, while before Hawking's discovery, one would have expected that, once formed, a black hole would be eternal, the fact that the radiation is carrying energy away, assuming overall energy conservation, leads one to expect that the mass of the black hole will start diminishing. The context in which this problem is standardly set is that of asymptotically flat spacetimes, for which we have a well defined notion of overall energy content given by the ADM mass³ of the spacetime, a quantity which is known to be conserved.

As we noted, Hawking's calculation cannot deal with back-reaction. However, our confidence on energy conservation in the appropriate situations is so robust that it is difficult not to conclude that, as the radiation carries away energy, the black hole mass will have to diminish. If this takes place, the surface gravity of the black hole—which is no longer really stationary, but can be expected to deviate from stationarity only to a very small degree—would change as well. As it turns out, the surface gravity is inversely proportional to the black hole's mass, so the black hole temperature can be expected to increase, leading to an ever more rapid rate of energy loss and a correspondingly faster decrease in mass.

The run away picture for the evaporation process suggests a complete disappearance of the black hole in a finite amount of time. Of course, we cannot really be sure about this picture because, in order to perform a solid analysis, we would need to deploy a, currently lacking, trustworthy theoretical formalism adept to the challenge. The

³The ADM mass is a quantity associated with the asymptotic behavior of the induced spatial metric of a Cauchy hypersurface. In asymptotically flat spacetimes, it is known to be independent of the hypersurface on which it is evaluated (see Arnowitt et al. (1962)).

problem is that, by the removal of energy from the black hole, one can expect to eventually reach a regime where quantum aspects of gravitation become essential to the description of the process. At such point, one might contemplate the possibility that, as a result of purely quantum gravitational aspects, the Hawking evaporation of the black hole will stop, leaving a small stable remnant. This, in turn, might open certain possibilities regarding the information issue. For the time being, though, we will ignore such an option.

Then, in order to simplify the discussion at this point, we will ignore the possibility of remnants and assume that there is nothing to stop the Hawking radiation. Then, if the black hole's mass decreases in accordance with energy conservation, one expects that the black hole to simply disappear and the spacetime region where it was located to turn flat (see Figure 2).

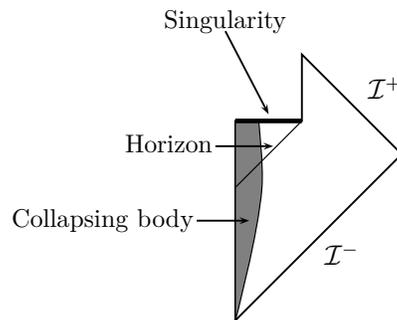


Figure 2: Penrose diagram for a collapsing spherical body, taking into account Hawking's radiation.

At this point, we seem to come face to face with an information loss problem: the original massive object that collapses, leading to the formation of a black hole, might have required an incredibly large amount of detail for its description. However, the final state that results from the evaporation is simply described in terms of the thermal Hawking flux, followed by an empty region of spacetime. More to the point, even if the initial matter that collapses to form a black hole was initially in a pure quantum state, after the complete evaporation of the black hole there would be a mixed one, corresponding to the thermal Hawking flux. These considerations seem to indicate that, even at the fundamental level, we have a fundamental loss of information. The final state, even if described in full detail, does not encode the information required to retrodict the details of the initial one. At the level of quantum theory, we would

be facing a non-unitary (and non-deterministic) relation between the initial and final states of the system, a situation that seems at odds with the unitary evolution provided by the Schrödinger equation.

There are, however, various caveats to the above conclusion. The first one is opened up by the possibility of the evaporation eventually stopping, leading to a stable remnant. The mass of said remnant can be estimated by considering the natural scales at which the effects of quantum gravity are expected to become important. This leads to an estimate of the order of Planck's mass ($\approx 10^{-5}$ gr). Then, if one wants the remnant to encode all the information present in the initial state, one is led to the conclusion that such a small object would have a number of possible internal states as large as that of the original matter that collapsed to form the black hole, which can, of course, have had a mass as large as one can imagine. It is hard, then, to envisage what kind of object, with such rather unusual thermodynamical behavior, would this remnant have to be. For this reason, this possibility is usually not considered viable (although we acknowledge that these considerations might be overturned; for a discussion of these issues see Banks (1994)). At any rate, we will not consider this possibility any further.

We should also mention another proposal which uses the idea that, while curing singularities, quantum gravity might open paths to other universes, which could be home to the missing information. Such information would be encoded either in a new universe or in correlations between it and ours. Besides the dramatic ontological burden, such proposal leaves open the possibility of these alternative universes emerging even in ordinary processes (which could, e.g., involve virtual black holes), leading to information loss in such standard scenarios. Alternatively, the information could be preserved, but impossible to retrieve in principle. We will also not consider this possibility any further.

A much more important caveat is the following: we have very solid results indicating that, associated with the formation of a black hole, there is always a singularity of spacetime appearing within it. The strongest results in this regard are a series of theorems proved by Hawking (see Hawking and Ellis (1973)) showing that, under quite general conditions, and assuming reasonable properties for the energy and momentum of the collapsing matter, the formation of singularities is an inevitable result of Einstein's equations. The issue is that, at the classical level, these singularities represent a breakdown of the theory and, in fact, a failure of the spacetime description. The singularities are, therefore, to be thought of as representing boundaries of spacetime, rather than points within it. Once a spacetime has additional boundaries, it is clear

that the issue of information has to be confronted on a different light. Of course, if one considers the description of the system at an initial Cauchy hypersurface and wants a final hypersurface to encode the same information, one has to make sure that the final one is also Cauchy.

The formation of singularities then implies that, if we want to have spacetime regions where the system's state could be thought of as encoding all the information, then we must surround the singularities by suitable boundaries. In other words, if the singularities force us to include further boundaries of spacetime, then the comparison of initial and final information has to be done between the initial Cauchy hypersurface and the late-time *collection* of surfaces that, together, act as a Cauchy hypersurface. That collection could naturally include asymptotically null future, but also the hypersurfaces surrounding the singularities. The same kind of calculation as the one done by Hawking would then show that all the information present on the initial hypersurface would also be encoded in the state associated with this late-time Cauchy hypersurface. That is, if we include the boundary of spacetime that arises in association with the singularity, then there is no issue regarding the fate of information. We conclude that, under these circumstances, still there is no information loss.

5 Second quantum gravity input: black holes do not involve singularities

As we noted above, singularities represent a breakdown of the spacetime description as provided by general relativity and thus indicate the need to go beyond such theory. The expectation among theorists is that quantum gravity is going to be the theory that cures these failures of classical general relativity, replacing the singularities by a description in the language appropriate to quantum gravity. This is, in fact, what occurs with various other theories that are known to be just effective descriptions of a physical system's behavior in a limited context, but that have to be replaced with a more fundamental description once the system leaves that regime. Think for instance of the description of a fluid by, say, the Navier-Stokes equations. We know that this description works very well in a large variety of circumstances, but that a breakdown of such description occurs, for instance, when there are shock waves or when other types of singularities are formed. However, under such circumstances, the underlying kinetic theory, including the complex inter-molecular forces, is expected to remain valid. The

point is that, just as in those cases, one expects the emergence of singularities in general relativity to indicate the end of the regime where the classical description of spacetime is valid and, therefore, where a quantum gravity description would have to take over (see Figure 3 and Ashtekar and Bojowald (2005) for details).

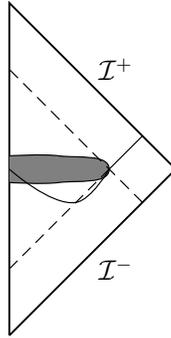


Figure 3: “Quantum spacetime diagram” for a black hole.

Of course, if quantum gravity does in fact cure the singularities, and removes the need to consider, in association with the corresponding regions, a boundary of spacetime, the issue of the fate of information in the Hawking evaporation of black holes resurfaces with dramatic force. So, do we finally have a genuine paradox in our hands. Not quite yet; a few elements are still missing. In order for a paradox to arise, we need to couple a genuine loss of information with a fundamental theory which does not allow for information to be lost.

6 A paradox?

When is it, then, that the Hawking radiation by a black hole leads to an actual paradox? We are finally in a position to enumerate the various assumptions required in order to construct a genuine conflict:

1. As a result of Hawking’s radiation carrying energy away from the black hole, the mass of the black hole decreases and it either evaporates completely or leaves a small remnant.
2. In the case where the black hole leaves a small remnant, the number of its internal degrees of freedom is bounded by its mass in such a way that these cannot possibly encode the information contained in an arbitrarily massive initial state.

3. Information is not transferred to a parallel universe.
4. As a result of quantum gravity effects, the internal singularities within black holes are cured and replaced by something that eliminates the need to consider internal boundaries of spacetime.
5. The outgoing radiation does not encode the initial information.
6. Quantum evolution is always unitary.

We have already discussed the arguments in support of assumptions 1, 2, 3 and 4 and saw that, although by no means conclusive, they are reasonable. But what about 5 and 6? Well, in order to avoid a paradox, and assuming the first four assumptions to be true, at least one of them has to be negated. In order to explore the motivations and consequences of doing so, we must think clearly about how to interpret Hawking’s calculation in a context in which 1, 2, 3 and 4 are the case.

As we remarked above, Hawking’s calculation is performed in the setting of a quantum field theory over a fixed curved background. What one finds there is that an initial pure state of the field evolves into a final one which, when tracing over the inside region, reduces to a mixed thermal state. The key question at this point, then, is how to interpret such a final mixed state in a setting in which i) the black hole is no longer there, so there is no interior region to trace over, and ii) in which there is no singularity (or parallel universe) for the information to “escape into.” As far as we can see, there are two alternatives: either one assumes that the mixed state arises only as a result of tracing over the interior region and maintains that the outgoing radiation somehow encodes the initial information—which amounts to negating 5; or one takes Hawking’s result seriously and maintains that, even in this scenario, information is lost—which amounts to negating 6. Below we explore each option in detail.

6.1 The outgoing radiation encodes information

In the last couple of decades, the community’s position on the information loss subject has been strongly influenced by developments in String theory. Such framework has permitted exploration of questions, regarding black holes, using settings where event horizons and singularities play no relevant roles. This is possible due to the AdS/CFT correspondence (see e.g., Strominger (2001)), which allows the mapping of complicated spacetime geometries in the “bulk” of asymptotically Anti-de Sitter spacetimes,

including ones involving black holes, onto corresponding states of an ordinary quantum field theory living on the Anti-de Sitter boundary (which is, in fact, a flat spacetime). These considerations have led people to conclude that, as a breakdown of unitarity is not expected to take place in the context of a quantum field theory in flat spacetimes, there should be no room for a breakdown of unitarity in the corresponding situation involving black holes either.⁴

The proposal, then, is that unitarity is never broken and that information is never lost. As a result, Hawking's calculation has to be somehow attuned to assure consistency. In particular, the proposal is that the outgoing radiation must encoded all of the initial information. There is, however, a high price to pay in order to achieve this. As has been shown in Almheiri et al. (2013), in order for the outgoing radiation to encode the necessary information, each emitted particle must get entangled with all the radiation emitted before it. However, due to the so-called, "monogamy of entanglement," doing so entails the release of an enormous amount of energy, turning the event horizon into a *firewall* that burns anything falling through it. The upshot then, is a divergence of the energy-momentum tensor of the field over the event horizon and a radical breakdown of the equivalence principle over such a region.

6.2 Unitarity is broken

The discovery of the Hawking radiation was initially taken as a clear indicative of information loss at the fundamental level. In fact, Hawking (1976) even introduced a notation for this general type of evolution which was supposed to account for the transformation from (possibly pure) initial states ρ_i into final mixed ones ρ_f . Hawking denoted the general linear, non-unitary, operator characterizing such transformation by the sign $\$$, i.e., $\rho_f = \$\rho_i$. Likewise, Penrose pointed out that, in order to have a consistent picture of phase space for situations involving black holes in thermal equilibrium with an environment, one has to assume that ordinary quantum systems undergo something akin to a self-measurement, by which he meant quantum state reduction that was not the result of measurement by external observers or measuring devices (see Penrose (1981)). Penrose (1999) further argued that quantum state reduction is probably linked to aspects of quantum gravity.

The early assessments of these ideas in Banks et al. (1984) indicated that they

⁴Note however that the argument can be easily reversed to show exactly the opposite. Since Hawking's result shows that unitarity breaks when black holes are present, one must conclude that quantum evolution *cannot* be unitary even in a quantum field theory on flat spacetimes.

where likely to lead to a very serious conflict with energy and momentum conservation or to generate unacceptable non-local features in ordinary physical situations. However, further analysis in Unruh and Wald (1995) showed that these assessments were not that solid and that there were various possibilities to evade the apparently damning conclusions.

In (omitted references) we have explored the viability of breaking unitarity both qualitatively and quantitatively. In particular, we have successfully adapted objective collapse models, developed in connection with foundational issues within quantum theory, in order to explicitly describe the transition from the initial pure state into a mixed one. Our view on the subject is based on the conviction that, contrary to the prevailing opinion in the community working on the gravity/quantum interface, there are good reasons to think that quantum theory requires modifications to deal with its basic conceptual difficulties. Below we discuss these issues and explore their consequences for the information loss paradox.

7 Information loss and the measurement problem

Most discussions of black holes and information loss do not implicate foundational issues of quantum theory. Of course, ignoring such issues, particularly with pragmatic interests in mind, is often acceptable. However, when deep conceptual questions are involved, such as in the present case, the pragmatic attitude might not be the right way to go.

The standard interpretation of quantum mechanics involves a profoundly *instrumentalist* character, with notions such as *observer* or *measurement* playing a crucial role. Such an instrumentalist trait becomes a problem as soon as one intends to regard the theory as a fundamental one, useful not only to make predictions in suitable experimental settings, but also to be applied to the measurement apparatuses, to the observers involved, or to non-standard contexts such as black holes or the universe as a whole. The resulting problem, often referred to as the *measurement problem*, has been discussed at length in numerous places and many different concrete formulations of it have been given. A particularly useful way to state it, given in Maudlin (1995), is as a list of three statements that cannot be all true at the same time:

- A. The physical description given by the quantum state is complete.
- B. Quantum evolution is always unitary.

C. Measurements always yield definite results.

Maudlin's formulation of the measurement problem is noteworthy because of its generality and its preciseness. Moreover, it is extremely useful in order to motivate and classify strategies to solve the problem. For example, by negating A, one arrives at so-called hidden variable theories, such as Bohmian mechanics; by removing B, one gets so-called objective collapse theories, such as GRW; and by discarding C, Everettian interpretations emerge. Of these three options, the last one is, by far, the most contentious. Among its most urgent matters, we can mention the problem of the preferred basis, the one of making sense of probabilities in the theory and the general and basic issue of establishing a clear and precise link between the abstract mathematical objects of the theory and concrete empirical predictions. Of course, brave attempts to deal with these and other issues within Everettian frameworks abound. However, we believe that, at least for the time being, they are far from being successful.

Returning to the measurement problem and its relation to the information loss issue, we note that assumptions 6 and B are in fact identical. Therefore, the strategy one decides to adopt in order to avoid complications regarding the information loss issue (e.g., negating 5 or 6 above) has implications with respect to what one must say regarding the measurement problem (e.g., negating A, B or C). In particular, if regarding the information loss, one decides to maintain the validity of 6 (and thus to hold that the outgoing radiation encodes all of the initial information), then one necessarily has to either negate A or C (i.e., either to entertain a hidden variables theory or an Everettian scenario). In other words, insisting on a purely unitary evolution, not only demands a violation of the equivalence principle and a divergence of the energy-momentum tensor, but also a commitment either with many worlds or with an acknowledgment that standard quantum mechanics is incomplete. On the other hand, if regarding the information loss problem, one decides to abandon unitarity, the same move automatically not only avoids a breakdown of the equivalence principle, but also guarantees success with respect to the measurement problem. The upper hand of the second option seems evident to us.

8 Conclusions

Since the publication of Hawking's analysis, more than forty years ago, the issue of black hole information loss has been a central topic in theoretical physics. The AdS/CFT

correspondence, proposed almost twenty years latter, came to further propel an already notorious debate. Yet, even after all these years, the discussion is often engulfed by confusion and misunderstanding among participants. The objective of this work is to develop a clear analysis of some of the key conceptual issues involved. Our hope is that, by doing so, significant progress on this important topic could soon be achieved.

We have presented the basic theoretical setting of the black hole information issue, paying special attention to elements, arising from not yet well-established physics, that presently have to be regarded merely as reasonable assumptions. Moreover, we have argued that the information loss issue is closely related to the measurement problem, and claimed that it is precisely within the context of certain proposals put forward to deal with the latter that the former finds one of its most conservative resolutions.

References

- Almheiri, A., Marolf, D., Polchinski, J., and Sully, J. (2013). Black holes: complementarity or firewalls? *JHEP*, 62.
- Arnowitt, R., Deser, S., and Misner, C. (1962). The dynamics of general relativity. In Witten, L., editor, *Gravitation: an introduction to current research*. Wiley.
- Ashtekar, A. and Bojowald, M. (2005). Black hole evaporation: a paradigm. *Class. Quant. Grav.*, 22(3349).
- Banks, T. (1994). Lectures on black hole information loss. *Nucl. Phys. Proc.*, 41.
- Banks, T., Susskind, L., and Preskin, M. E. (1984). Difficulties for the evolution of pure states unto mixed states. *Nucl. Phys. B*, 224(125).
- Bekenstein, J. D. (1972). Black holes and the second law. *Lett. Nuovo Cim.*, 4(737).
- Hawking, S. W. (1976). Breakdown of predictability in gravitational collapse. *Phys. Rev. D*, 14(2460).
- Hawking, S. W. and Ellis, G. F. R. (1973). *The large scale structure of spacetime*. Cambridge University Press.
- Maudlin, T. (1995). Three measurement problems. *Topoi*, 14.

Penrose, R. (1981). Time asymmetry and quantum gravity. In Isham, C. J., Penrose, R., and Sciama, D. W., editors, *Quantum Gravity II*. Clarendon Press.

Penrose, R. (1999). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

Strominger, A. (2001). The AdS/CFT correspondence. *JHEP*, 0110(034).

Unruh, W. G. and Wald, R. M. (1995). On evolution laws taking pure states to mixed states in quantum field theory. *Phys. Rev. D*, 52:2176–2182.

Wald, R. M. (1994). *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*. University of Chicago Press.

The Causal Homology Concept

Jun Otsuka*

Abstract

This presentation proposes a new account of homology, which defines homology as a correspondence of developmental or behavioral mechanisms due to common ancestry. The idea is formally presented as isomorphism of causal graphs over lineages. The formal treatment not only clears the metaphysical skepticism regarding the homology thinking, but also provides a theoretical underpinning to the concepts like constraints, evolvability, and novelty. The novel interpretation of homology suggests a general perspective that accommodates evolutionary developmental biology (Evo-Devo) and traditional population genetics as distinct but complementary approaches to understand evolution, facilitating further empirical and theoretical researches.

*Department of Philosophy, Kobe University, Rokko-dai 1-1 Nada, Kobe, Japan. Email: junotk@gmail.com

1 Introduction

The homology thinking, the idea that the same anatomical structure repeatedly appears in different species or parts of the same organism, has a long history in biology (Amundson, 2005). While the existence of such anatomical similarities among or within species is now explained by the descent from a common ancestor, the conceptual issues surrounding the notion have invited philosophical as well as methodological debates and skepticism. Owen famously defined homology as “the same organ in different animals under every variety of form and function,” but this definition is perplexing rather than enlightening: what characterizes and warrants the sameness of “organs,” if not their form or function? What, in other words, is the unit of homology?

There are three conceptual problems. The first and foremost problem is its *definition*: what exactly is homology? Evolutionary theory tells us that homology is identity due to a common origin, but an identity of *what*? Is it morphological characters, activities, clusters of properties, or genetic networks that are regarded to be same? And what is the criterion to judge whether or not two such things are actually the “same”? The second problem is *metaphysical*. As Ghiselin (1997) points out, the homology-as-identity partitions the whole tree of life into equivalence classes. But doesn’t the supposition of such universal classes, reminiscent of Aristotelian essence, commit us to an anti-evolutionary thinking? And thirdly, there is a *pragmatic* question: why do we care about homology at all? Some neo-Darwinians such

as G. C. Williams see homologs as mere “residues,” i.e. a relic of the past common ancestry not yet washed out by natural selection (Amundson, 2005, pp. 237-8). If that is the case homology by itself would have no explanatory role in evolutionary theory, and the quest for its definition, however well-defined and metaphysically sound, becomes a mere armchair exercise with no scientific value.

There is at least one usage of the concept free from these issues: homology of DNA sequences. Here the “sameness” is well-defined by matching bases that can be one of the four chemical kinds, G, C, T, A. Moreover, the scientific importance of orthologs and paralogs is undeniable in reconstructing the evolutionary history and predicting gene function, to name a few. Things become different for phenotype, in particular complex phenotypes like morphological or behavioral traits. First of all, there is no clear-cut definition of “phenotypic units” as that for nucleotides. Continuous traits such as height or weight usually lack objects breakpoints by which we classify them into discrete equivalence classes. In sum, there seem to be no non-arbitrary and non-controversial units for phenotype of which we can talk about the sameness, and thus homology.

Our first task, therefore, is to identify the units on which the phenotypic homology relationship can be defined. This presentation proposes that this purpose is best served by *causal graphs* which formally represent developmental or behavioral mechanisms. Homology is thus defined as graph isomorphism over lineages, or conservation of the underlying causal structure

over evolutionary history (Section 2). I will argue in Section 3 that the formal treatment of homology (i) solves the philosophical as well as empirical puzzles and criticisms regarding the homology concept; (ii) provides clear meanings to some key but elusive concepts such as constraints, evolvability, and novelty; (iii) and suggests a broad perspective that accommodates evolutionary developmental biology (Evo-Devo) and traditional population genetics as distinct but complementary research projects. Section 4 compares the present approach to other existing accounts of homology, and discusses its relative strengths, challenge, and philosophical implication. As will be stressed there, the primary objective of this presentation is to facilitate or open up new empirical as well as theoretical questions. The last section concludes with some of these research prospects that are prompted by the new homology concept.

2 Defining homology with graphs

The idea of characterizing homology in terms of causal structures is not new. Various biologists have suggested, albeit in different fashions, that the developmental or behavioral mechanisms underlying phenotype can or should serve as a unit of homology (e.g. Riedl, 1978; Wagner, 1989, 2014; Gilbert and Bolker, 2001; Müller, 2003). These proposal, however, are mostly based on independent examples or qualitative descriptions, and the lack of a unified treatment has blurred their philosophical as well as theoretical implications.

The aim of this section is to give a formal representation to the ideas of developmental sameness by using causal graphs, in view of exploring the conceptual nature of homology in the later sections.

A *causal graph* \mathcal{G} is a pair (\mathbf{V}, \mathbf{E}) , where \mathbf{V} is a set of phenotypic or genetic variables of organisms and \mathbf{E} is a set of edges representing causal relationships among these traits. Development is understood as a causal web connecting embryological, morphological, and behavioral traits, and the set of edges \mathbf{E} characterizes these causal links. Note that such connections may remain invariant even under considerable modifications in phenotypic values or the functional form that determines the quantitative nature of each edge. The same set of \mathbf{E} is consistent with a variety of phenotypic states and forms of causal production; it only defines the qualitative feature of the causal networks, i.e. which causes which.

Once modeled in this way, it becomes meaningful to compare causal structures of different organisms. A causal graph $\mathcal{G}_1 = (\mathbf{V}_1, \mathbf{E}_1)$ is *isomorphic* to another $\mathcal{G}_2 = (\mathbf{V}_2, \mathbf{E}_2)$ if they have the same structure, or more formally if there is a bijection $f : \mathbf{V}_1 \rightarrow \mathbf{V}_2$ such that if $(v, w) \in \mathbf{E}_1$ then $(f(v), f(w)) \in \mathbf{E}_2$. Likewise, isomorphism can be defined for subgraphs, which are just parts of the causal graphs restricted to a subset $\mathbf{V}' \subset \mathbf{V}$. We write $\mathcal{G}_1 \sim \mathcal{G}_2$ if two (sub)graphs are isomorphic. It is easy to see ‘ \sim ’ is symmetric, reflexive, and transitive, and thus defines a equivalence class.

Each individual is assigned one causal graph that models a particular part of its developmental or behavioral mechanism. Let us denote the causal

structure of an organism a by $\mathcal{G}(a)$. Collectively, $\mathcal{G}(A)$ is a set of causal structures for a set of organisms A . We assume usual ancestor/descendant relationships over a set of organism Ω (which may include more than one species). If b is an ancestor of a , the *lineage* between b and a is a set of every individual between them. Given this setup homology is defined as follows.

For two sets of organisms $A, B \subset \Omega$, let \mathcal{G}' be a subgraph of all $g \in \mathcal{G}(A)$, and \mathcal{G}'' be a subgraph of all $g \in \mathcal{G}(B)$. Then \mathcal{G}' and \mathcal{G}'' are homologous iff

1. $\mathcal{G}' \sim \mathcal{G}''$;
2. there is a set of common ancestors $C \subset \Omega$ of A and B ¹; and
3. for every d in all the lineages from C to A and C to B , $\mathcal{G}(d)$ has a subgraph \mathcal{G}''' such that $\mathcal{G}''' \sim \mathcal{G}' \sim \mathcal{G}''$.

The definition explicates the idea that homology is the identity between causal structures due to common ancestry. Two (sets of) organisms share a homologous causal structure if, in addition to the graph isomorphism, every individual on the lineage connecting them shares the same causal graph, capturing the idea that the structure has been conserved through the evolutionary history.

The same treatment applies to serial homology, i.e. the homology relationship among parts of the same organism, such as teeth, limbs, or tree

¹Note that C may be A or B themselves. Also note the condition 1 is redundant if a lineage includes the both ends. But here it is retained for clarity.

leaves. We can just set $A = B$, and compare different but isomorphic subparts $\mathcal{G}', \mathcal{G}''$ of the same overall structure $\mathcal{G}(A)$. Then the homology hypothesis is that there is an organism c in which the mechanism in question was duplicated, and the lineages from c to A have conserved the duplicated structures.

The above definition is illustrated with a case of special homology in figure 1, which depicts a particular region of the tree of life for (groups of) organisms A to G . Two mutations M_1, M_2 on the developmental mechanism occurred in the lineage leading to F , in which one causal edge $V_1 \rightarrow V_3$ was first removed and then restored. In this example, the causal structure $\mathcal{G}(D)$ of population D is homologous to $\mathcal{G}(E)$, for they are both inherited from the ancestral graph $\mathcal{G}(B)$ and $\mathcal{G}(A)$. In contrast, it is not homologous to $\mathcal{G}(F)$ even though they are graph-isomorphic. This is because the lineages connecting D and F do not conserve the causal structure in question: particularly it is not shared by C .

The example, though too simplistic to capture any real biological phenomena, makes explicit the idea that homology is a concordance of developmental mechanisms due to common ancestry. Note the criterion makes no reference to the resulting phenotype represented by particular values or distributions of variables. It does not require or forbid that, for example, two populations E and D show similar morphological distributions. Nor does it assume the graphs consist of the variables of the same nature. If the causal graphs in figure 1 represent a genetic network, kinds of genes/variables that

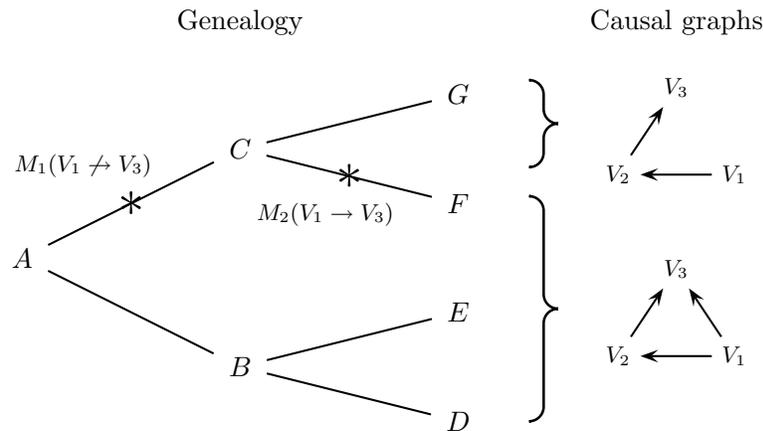


Figure 1: Illustration of graph homology. On the left is a genealogy tree for hypothetical populations A, B, C, D, E, F, G , while the graphs on the right describe causal structures of these populations over three characters, V_1, V_2 , and V_3 . Two asterisks (*) on the tree denote mutation events on the causal structure. See text for explanation.

constitute the network may vary across populations, as long as they serve the same causal roles within the overall structure. It is structural, rather than material, identity that defines homology. Theoretical as well as philosophical implications of this view will be explored in the following sections.

3 Conceptual advantages of the view

The above account is intended to provide a theoretical platform to formulate and evaluate hypotheses or explanations regarding homology. This section explicates the conceptual benefits of thinking homology in terms of causal graphs. Discussions on the empirical adequacy are deferred to the next sec-

tion.

As discussed in the introduction, the major obstacle in defining homology is the absence of definite phenotypic units. Homology is an identity rather than similarity relationship (e.g. Ghiselin, 1997; Müller, 2003; Wagner, 2014), whereas no two or more phenotypic characters are identical in a strict sense — there are always subtle differences in, say, shape or size. The problem could be solved if we could find a natural and non-arbitrary way to factorize the phenotypic space into discrete regions so that two phenotypes within the same region are regarded “identical” despite their apparent differences. This is a difficult task, especially because we do not know the topological feature of the phenotypic space (Wagner and Stadler, 2003). To solve this issue the present analysis adopts a different strategy: instead of trying to impose a certain structure on the phenotypic space, it takes the generative mechanisms as basic units. Once these mechanisms are represented by causal graphs, which by nature are discrete mathematical entities, the desired identity relationship is given by graph isomorphism regardless of differences in the resulting morphology/phenotype. The graphical representation thus provides natural units prerequisite to define homology.

It is granted that a graph representation is not determined uniquely, because the same developmental mechanism can be modeled in various levels of abstraction, yielding causal graphs of different complexities. However, I take this to be a strength rather than weakness of my view, because homology too is often treated as description-dependent. Teleost fins and tetrapod limbs

are said to be homologous *as* paired vertebrate appendages, but *not as* fins or limbs. In contrast, our hands and pectoral fins of the whale are homologous not only as appendages but also as limbs. One tempting hypothesis is that such degrees of homology relationship correspond to isomorphisms of causal structures described at different granularities. In the above example, it is hypothesized that teleost fins and tetrapod limbs are represented by the same, but rather course-grained, causal graph, while tetrapod species share the causal structure to much finer details.

Fixing the level of abstraction determines not only the equivalent classes but also the degree of similarity between these classes. Two distinct causal graphs may be closer or further depending on the number of changes required to obtain one from the other. If \mathcal{G}'' is obtained by removing one edge from \mathcal{G}' which in turn lacks one of the edges of \mathcal{G} , \mathcal{G}'' is one step further than \mathcal{G}' from the original \mathcal{G} . Each such deletion or addition of causal connection is called *novelty*. Novelty in this framework is a modification of the causal graph, and as such creates a new equivalence class of causal graphs, namely homology. Evolutionary novelty also comes in different degrees. In general, a single modification in abstract graphs will correspond to multiple edge additions or deletions in detailed ones, and thus is weighted more. In this regard a change in the causal graph shared both by teleosts and tetrapods will count as a significant novelty and possibly a creation of a new “bauplan.”

This brings us to one of the central contentions in today’s evolutionary biology, namely the alleged inadequacy of the Modern Synthesis framework,

in particular population genetics, to incorporate macro-scale evolutionary phenomena uncovered by evolutionary developmental biology (e.g. Pigliucci and Müller, 2010). It has been claimed that homology (macro-scale conservatism) and novelty (a large phenotypic change) not only resist explanations by the Neo-Darwinian gradualism, but also constrain evolutionary trajectories as modeled in population genetics (e.g. Amundson, 2005; Brigandt, 2007). The theoretical relationship between Evo-Devo and population genetics, however, remains elusive, which makes difficult to evaluate the call for the “new synthesis.”

The present approach, by expressing homology and novelty in terms of graph equivalence and modification, suggests a perspective on this connection and a way to turn these claims into empirical hypotheses. Because causal models induce evolutionary changes as studied in population and quantitative genetics (Otsuka, 2015, 2016), the graphical representation allows one to analyze how developmental structures generate and constrain evolutionary dynamics. In particular, topological features of the graph such as modularity yield, via the so-called Markov condition, patterns of probabilistic independence on the phenotypic distribution and determine possible evolutionary trajectories or *evolvability*. The causal graph approach thus supports the view that a homolog constitutes a unit of morphological evolvability (Brigandt, 2007).

The graph structures that yield population dynamics are usually not study objects of population genetics. They rather serve as background frame-

works in which evolutionary models are build to study changes in genetic or phenotypic frequencies. These frameworks, however, must come from somewhere, and this evolutionary process is a primary interest of Evo-Devo. Studies on homology and novelty — graph stasis and change — amount to “higher order” evolutionary analyses that deal with changes in the theoretical framework used in population genetics to predict local population dynamics. The graphical conception of homology thus suggests a broad perspective that accommodates these different, and sometimes seen antagonistic, research fields as complementary approaches to understand evolution.

Finally, let us turn to the metaphysical problem. As seen above, homology is defined as an equivalence class over a set of causal graphs. But to what do such classes correspond, if not some ideal types or essences? Homology thinking has been criticized as anti-evolutionary due to its alleged commitment to essentialism. These critics thus re-interpret homology as a lineage that connects individual parts, rather than as a universal class to be instantiated by its members/homologs (e.g. Ghiselin, 1997). A detailed examination of this criticism must await another occasion, but here I just want to propose a different way to look at the issue. A metaphysical implication from the present study is that homology stands to concrete parts of organisms not as a universal to individuals, nor as a whole to parts, but rather as a model to phenomena to be modeled. A homology hypothesis is based on an observation that two or more individuals or parts thereof can be modeled by

the same causal graph.² Hence the proper relationship is not instantiation or mereology, but representation (Suppes, 2002). Once conceived in this way, the metaphysical ghost of essentialism vanishes away. Just like the same oscillator model characterizes various kinds of pendulum clocks, homology-as-model is a mathematical entity (directed graph) that may represent more than one actual individual, but that does not force us to commit to any form of essentialism.

The individual-universal distinction has also cast a shadow on the pragmatic issue regarding the epistemic role and significance of the concept of homology. It has been argued that the study of homology cannot be any more than a historiography since there is no such thing as a law for individuals (Ghiselin, 1997). A very different picture, however, emerges from the present thesis. A homology statement is a historical hypothesis regarding causal isomorphism — that two or more (sets of) organismal parts can be represented by the same causal model — and as such makes various predictions. For example, it supports extrapolations from model organisms, predicting that homologous organs will respond in the same or similar fashion to physiological, chemical, or genetic interventions. In addition, since isomorphic developmental structures will generate similar patterns of phenotypic variation (see above), their evolutionary changes are expected to follow similar trajectories. Establishing homologous relationships therefore is not a mere

²This, in turn, implies these individuals would respond in a more or less same fashion to hypothetical interventions (Woodward, 2003). Hence homology statements eventually boil down to counterfactual claims.

historical description, but has predictive implications both on physiological and evolutionary studies.

4 Comparisons and possible objections

This section compares the present proposal with some of the existing accounts of homology and also discusses possible objections. A number of philosophers and biologists have recently proposed to define homology as a *homeostatic property cluster*, a cluster of correlated properties maintained by “homeostatic mechanisms” (e.g. Boyd, 1991; Rieppel, 2005; Brigandt, 2009; Love, 2009). Since clustering and correlations are a matter of degree, homology according to this view is not an identity but a similarity relationship. It thus confronts with the boundary problem — to what extent properties must be clustered to form a homolog? The underlying “homeostatic mechanism” is supposed to clarify this boundary, but without a clear definition of what it is such an attempt only leads to a circularity. In particular, if it is defined as “those causal processes that determine the boundary and integrity of the kind (Brigandt, 2009, p.82),” the charge of circularity cannot be avoided.

This kind of problem will not arise if the generative mechanisms are defined explicitly in terms of causal graphs. While my approach proposes a formal framework to represent these mechanisms, it does not make any assumption or restriction on their structure: in particular it does not require the mechanism to be homeostatic, circumventing the criticism that a home-

ostatic mechanism by definition cannot evolve (Kluge, 2003). Moreover, the reference to “clusters” or even properties becomes superfluous, because the variational properties of phenotype are mere derivatives of the underlying causal graph. Of course, covarying traits suggest some ontogenetic connections, and thus may serve as a useful heuristics for finding homologs. They are, however, only “symptoms” — what *define* homology are not properties, clustered or homeostatic, but rather generative mechanisms.

The present approach has a closer affinity to the so-called *biological homology concept* that attempts to explain the phenomena of homology on the basis of a particular feature of the underlying causal structure, such as gene regulatory networks (e.g. Wagner, 1989, 2014). Indeed, one motivation of this presentation is to give a formal platform for these empirical hypotheses to elucidate their theoretical as well as philosophical implications. An important empirical challenge to the biological homology concept, and any other attempts to identify a homolog with a certain developmental structure, is the well-known fact that morphological similarity does not entail developmental sameness (Wagner and Misof, 1993). It has been reported that apparently homologous characters in related species may develop from different genes, cell populations, or pathways — the phenomena called *developmental system drift* (True and Haag, 2001). Although these phenomena present a challenge to my account as well, not all of them count as counter evidence. If, for example, “drift” concerns only genetic or cell materials, topological features of the causal network may remain invariant. Descriptive levels also matter.

Even if two causal structures differ at a fine-grained description, they may coincide at a more abstract level. Finally, my view does not require the entire developmental system to be conserved: if causal graphs share *some* part, they may still be homologous *in that aspect*. Indeed, it would be surprising if two apparent homologs turn out to share no developmental underpinnings at all. Some degree of flexibility may be expected, but so is inflexibility. Representing and comparing homologs in terms of the underlying causal graphs will serve as a heuristics to identify which part of the overall developmental system is responsible for generating similar morphological patterns.

From a philosophical perspective, a distinguishing feature of my account is its explicit reference to *models*. Homology has traditionally considered to be a relationship among concrete biological entities or properties thereof: it is organs or phenotypic features that are said to be homologous. In contrast, homology in my view is a relationship among abstract entities, i.e. causal graphs. How and why does such an abstract relationship reveal anything interesting about the concrete evolutionary history? That scientific theories and concepts should directly describe actual phenomena is a predominant view of science both in lay and scholarly circles. Under this conception logical positivists made it their primary task to define theoretical terms by the observable. In the same vein philosophers of biology have tried (not successfully in my view) to justify the concepts like homology or species by identifying necessary and sufficient conditions in terms of visible or directly verifiable features of organisms.

This apparently intuitive picture, however, has been criticized to be an overly simplistic view on the relationship between a scientific theory and reality (e.g. Suppes, 1967; Cartwright, 1983; Suppe, 1989). According to the critics the primary referents of scientific theories, concepts, and laws are not actual phenomena but idealized models. These models are not exact replicas of reality, but extract only certain features that are supposed to play essential roles in the scientific problem at hand. The present analysis is in line with this tradition. Causal graphs are highly idealized and thus possibly incomplete representations of complex causal interactions in living systems, but it is this idealization that affords explanatory power and general applicability. That is, on the condition that a model extracts the common causal structure of a population can it be used to predict the population's evolutionary trajectory or consequences of hypothetical interventions.

Most of these models, however, are still idiosyncratic to particular populations — e.g. population geneticists usually build, customize, or parameterize their model for each study object.³ Homology thinking aims at even higher generality: its core idea is that some distinct species or organs allow for the same treatment/model in the analyses of their evolutionary fate or physiological performance. A homology statement is a historical hypothesis as to why such a unified explanation is possible at all. That is, it justifies the use of the same causal model based on evolutionary history, i.e. by the descent of the

³Models of adaptive evolution, however, may be extrapolated to the same or similar environmental conditions. In this regard, the analogical thinking and homological thinking represent two distinct ways to generalize evolutionary models.

causal graph from common ancestry. Hence homology is far from “residual,” but has a significant explanatory value in biology — it allows an extrapolation of an evolutionary or physiological model to other contexts, and thus provides a basis for the highest-level generality in biological sciences.

5 Conclusion

The concept of homology presupposes phenotypic units on which identity relationships can be defined. The present analysis identified these units with causal graphs representing developmental or behavioral mechanisms and defined homology as graph isomorphism over lineages. The advantage of this formal concept is that it acknowledges the distinctive role of the study of homology while suggesting its connection to the traditional population genetics framework. That is, it not only provides definite meanings to such concepts like constraints, evolvability, and novelty, but also presents homology as a historical account or justification of the generalizability of evolutionary or physiological models. This is paralleled with the shift in the ontological nature of what can be said to be homologous: homology is a relationship between theoretical models, rather than concrete biological entities such as organs. Hence the proper relationship between homology to actual biological phenomena is not instantiation, but representation. Once conceived in this way the metaphysical problem of the alleged essentialism fades away.

The new account of homology prompts empirical, theoretical, and philo-

sophical researches on various topics, including the study of novelty and evolvability, the interplay between Evo-Devo and population genetics, implications of developmental flexibility, and the generalizability of biological models, to name a few. Another interesting philosophical question not mentioned above is the possibility of extending the current approach to another vexing concept in evolutionary biology, namely *species*. If homology is a partial matching of the causal structures between distinct species, it is tempting to define species by the whole causal structure — so that two organisms belong to the same species if their entire ontogeny and life history are represented by the same causal graph. This is a big question that requires an independent analysis, but will be briefly discussed in the presentation if time permitted.

References

- Amundson, R. (2005). *The Changing Role of the Embryo in Evolutionary Thought: Roots of Evo-Devo*. Cambridge University Press, New York, NY.
- Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, 61(1-2):127–148.
- Brigandt, I. (2007). Typology now: homology and developmental constraints explain evolvability. *Biology & Philosophy*, 22(5):709–725.

- Brigandt, I. (2009). Natural kinds in evolution and systematics: Metaphysical and epistemological considerations. *Acta Biotheoretica*, 57(1-2):77–97.
- Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press, New York, NY.
- Ghiselin, M. (1997). *Metaphysics and the Origin of Species*. State University of New York Press, New York.
- Gilbert, S. F. and Bolker, J. A. (2001). Homologies of process and modular elements of embryonic construction. *Journal of Experimental Zoology*, 291(1):1–12.
- Kluge, A. G. (2003). On the deduction of species relationships: A précis. *Cladistics*, 19(3):233–239.
- Love, A. C. (2009). Typology reconfigured: From the metaphysics of essentialism to the epistemology of representation. *Acta Biotheoretica*, 57(1-2):51–75.
- Müller, G. B. (2003). Homology: The Evolution of Morphological Organization. In Müller, G. B. and Newman, S. (eds.), *Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology*, pp. 51–69. The MIT Press.
- Otsuka, J. (2015). Using Causal Models to Integrate Proximate and Ultimate Causation. *Biology & Philosophy*, 30(1):19–37.

- Otsuka, J. (2016). Causal Foundations of Evolutionary Genetics. *The British Journal for the Philosophy of Science*, 67(1): 247-269.
- Pigliucci, M. and Müller, G. B. (2010). *Evolution: the extended synthesis*. MIT Press, Cambridge, MA.
- Riedl, R. (1978). *Order in living organisms: a systems analysis of evolution*. Wiley, New York, NY.
- Rieppel, O. (2005). Modules, kinds, and homology. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 304(1):18–27.
- Suppe, F. (1989). *The semantic conception of theories and scientific realism*. University of Illinois Press.
- Suppes, P. (1967). What is a scientific theory? In Morgenbesser, S. (ed.), *Philosophy of Science Today*, pp. 55–67. Basic Books, Inc., New York.
- Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. CSLI Publication, Stanford, CA.
- True, J. R. and Haag, E. S. (2001). Developmental system drift and flexibility in evolutionary trajectories. *Evolution and Development*, 3(2):109–119.
- Wagner, G. P. (1989). The biological homology concept. *Annu. Rev. Ecol. Evol. Syst.*, 20:51–69.
- Wagner, G. P. (2014). *Homology, Genes, and Evolutionary Innovation*. Princeton University Press, Princeton, NJ.

Wagner, G. P. and Misof, B. Y. (1993). How can a character be developmentally constrained despite variation in developmental pathways? *Journal of Evolutionary Biology*, 6(3):449–455.

Wagner, G. P. and Stadler, P. F. (2003). Quasi-independence, homology and the unity of type: a topological theory of characters. *Journal of theoretical biology*, 220(4):505–527.

Woodward, J. B. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York.

Serendipity: an argument for scientific freedom?

Baptiste Bedessem and Stéphanie Ruphy
Université Grenoble Alpes

PSA 2016, Atlanta

Abstract

The unpredictability of the development and results of a research program is often invoked in favor of a free, disinterested science that would be led mainly by scientific curiosity, in contrast with a use-inspired science led by definite practical expectations. This paper will challenge a crucial but underexamined assumption in this line of defense of scientific freedom, namely that a free science is the best system of science to generate unexpected results. We will propose conditions favoring the occurrence of unexpected facts in the course of a scientific investigation and then establish that use-inspired science actually scores better in this area.

1. Introduction

“I didn’t start my research thinking that I will increase the storage capacity of hard drives. The final landscape is never visible from the starting point.” This statement made by the physicist Albert Fert (2007), winner of the 2007 Noble Prize for his work on the giant magnetoresistance effect, expresses a very common belief, especially among scientists, about the unpredictable nature of the development and results of a research program. Such retrospective observations feed a type of “unpredictability argument” often invoked in favor of a pure, disinterested science led by scientific curiosity, in contrast with a use-inspired or applied science led by practical considerations. Polanyi gave a somewhat lyrical form of this kind of unpredictability argument in his classical essay “The Republic of Science” (1962). Science, says Polanyi (1962, 62), “can advance only by unpredictable steps, pursuing

problems of its own, and the practical benefits of these advances will be incidental and hence doubly unpredictable. ... Any attempt at guiding research towards a purpose other than its own is an attempt to deflect it from the advancement of science... You can kill or mutilate the advance of science, but you cannot shape it.” In Polanyi’s view, claims about the unpredictable nature of scientific development go hand in hand with a plea for an *internal* definition of research priorities: a problem should be considered important in light of considerations internal to a field of scientific inquiry and not (at least not primarily) in light of external considerations, such as practical utility. The orientation of the inquiry by practical objectives is then deemed epistemically counter-productive and vain: one should not attempt to predict the unpredictable.

In response to this line of defense of free science, some authors emphasize the epistemic fecundity of use-inspired science (Stokes 1997, Wilholt 2006, Carrier 2004) showing that the presence of practical objectives does not run counter to the building of fundamental knowledge: more fundamental knowledge may be needed to achieve some particular practical ends. Industry research on the giant magnetoresistance effect in the 1990s is a telling example of research undertaken under considerable pressure to produce applicable results but which nevertheless produced, along the way, new fundamental knowledge (Wilholt 2006).

Our aim in this paper is to develop another line of defense of the epistemic fecundity of applied science, by challenging a crucial but often implicit assumption in the traditional defense of scientific freedom based on scientific unpredictability (such as Polanyi’s or Fert’s), namely the assumption that a free science is the best system of science to generate unexpected facts. But what are actually the conditions favoring the emergence of novelty in the course of a scientific investigation? This important issue has not received much epistemological

attention.¹ We will fill this gap by first distinguishing two kinds of unpredictability arguments often mixed when debating on scientific freedom, to wit, unpredictability as unforeseen practical applications and unpredictability as *serendipity* (cases, as we will explain in more details, where unexpected facts open up new lines of inquiry). Focusing on the latter, we will propose two conditions that favor the occurrence of unexpected facts in the course of a scientific investigation. In light of these two criteria we will then compare pure, disinterested science and applied science as regards their capacity to generate novelty.

2. Two types of unpredictability arguments

Appeals to the unpredictability of scientific results actually refer to various kinds of situations, which need to be clearly distinguished. First, the notion of unpredictability of scientific results can designate unforeseen practical applications of fundamental knowledge. Second, it can refer to a serendipitous dynamics of scientific progress: a line of research may sometimes lead to a totally unexpected, surprising result, which opens a new direction of inquiry. These two kinds of unpredictability give rise to *distinct* arguments in favor of scientific freedom, unfortunately often mixed in discussions about the relative merits of pure science and application oriented science.

2.1 Unpredictability as unforeseen practical applications

When unpredictability refers to unexpected applications, the argument is the following: freedom of research should be preserved since a free, disinterested science is needed to generate a reservoir of fundamental knowledge, which then can be used to develop

¹ Wilholt and Glimell (2011, 353) do touch upon this issue when discussing the link made by proponents of the autonomy of science between freedom of research and diversity of approaches favoring the epistemic productivity of science. But they just note that it is a strong assumption and do no further discuss its validity.

applications. This argument was typically developed by Vannevar Bush who appealed to the now classically called linear model of innovation:

“Basic research leads to new knowledge. It provides scientific capital. It creates the fund from which the practical applications of knowledge must be drawn. New products and new processes do not appear full-grown" (1945, 20).

The development of the H-bomb in the frame of the Manhattan project is a paradigmatic case, also invoked by Bush: “basic discoveries of European scientists" (1945, 20) about the structure of the matter is what made possible the military application. Another frequently cited example of unpredictable application is the invention of the laser, a widely-used technological device nowadays, made possible by pure theoretical developments in quantum physics during the first half of the XXth century.

We will not in this paper discuss further this first version of the unpredictability argument. Let us just mention that its underlying linear model of innovation linking pure science and practical applications has already been challenged on several grounds by various authors (e.g. Brooks, 1994; Leydesdorff, 1997; Edgerton, 2004; Rosenberg, 1992). We rather want to focus on the second (and also widespread) type of unpredictability arguments, whose validity has been much less scrutinized.

2.2 *Unpredictability as serendipity*

This second type of argument appeals to unpredictability in the sense of *serendipity*: an unexpected observation or result opens up a new line of research leading to a fundamental discovery. A very well known historical episode illustrating such a serendipitous scientific dynamics is the invention of the first antibiotic by Flemings, after he had accidentally

observed the effect of a fungi (*Penicilium*) on bacteria colonies (Flemings, 1929). Also often cited is the discovery of radioactivity by Henri Becquerel (1896): when working with a crystal containing uranium, Becquerel noted that the crystal had fogged a photographic plate that he had inadvertently left next to the mineral. This observation led to the hypothesis that uranium emitted its own radiations. Another, perhaps less cited instance of serendipitous scientific dynamics is the discovery of the chemotherapeutic cisplatin molecule by scientists initially working on the effects of an electric field on bacteria growth (Rosenberg *et al.*, 1967). They observed that cell division was inhibited because of the unexpected formation of a chemical compound with the Platinum atoms contained in the electrode. This chemical compound, which they named cisplatin, was then successfully tested as an anti-proliferative agent against tumoral cells.

When unpredictability refers to such serendipitous discoveries, freedom of research is defended on the grounds that scientists should be able to freely change the direction of their research or open up new lines of inquiry, in order to be able to follow up on unexpected results, thereby generating new knowledge (which in turn will possibly lead to new applications). But to properly work as an argument favoring free, disinterested research over applied research, this “serendipity argument” actually presupposes that the occurrence of surprising facts is more likely to happen in the first system of science than in the second. For increasing the production of new knowledge (and possibly new applications) does not only depend on being able to freely follow up on unexpected facts, it also (obviously) depends on whether occurrences of unexpected facts are favored, to start with. Two types of considerations are thus mixed in the serendipity argument: considerations on the occurrence of unexpected facts and considerations on the (institutional, material) possibility to follow up on them.

We will not for the moment discuss the second type of considerations and focus on the first, which has been largely neglected in the literature on scientific freedom, namely the conditions that favor the occurrence of surprising facts. Our central issue is thus the following: is a use-inspired science less likely to generate unexpected results than a free science mainly fuelled by curiosity? After having clarified the notion of *unexpected* result, we will propose two criteria that, we will argue, favor the occurrence of such results and in light of which free science and applied science can be compared.

3. Conditions of emergence of unexpected facts

By “unexpected facts” occurring in the course of an inquiry, we simply mean here results (observations, outcomes of an experiment, etc.) that cannot be accounted for within the theoretical or, more largely, the epistemic framework in which the empirical inquiry has been conceived and conducted. This kind of “exteriority” is what leads scientists to move away from the initial explanatory framework and open up new lines of inquiry in search of an alternative one that could accommodate the unexpected results.

3.1 Isolation and purification of phenomena

It is now a well-known feature of contemporary experimental sciences that many of their objects under study are “created” in the laboratory rather than existing “as such” in the real world. When drawing our attention to this epistemologically important feature, Hacking (e.g. 1983, chap. 13) specified that we should not read this notion of “creation” of phenomena as if *we were making* the phenomenon, suggesting instead that a phenomenon is “created” in the laboratory to the extent that it does not exist outside of certain kinds of apparatus. This is typically the case for a phenomenon like the Hall effect: it did not exist “until, with great ingenuity, [Hall] had discovered how to *isolate, purify* it, create it in the laboratory” (Hacking

1983, 226, *our italics*). In other words, Hall created in 1879 the material arrangement - a current passing through a conductor, at right angles to a magnetic field -, for the effect to occur and “if anywhere in nature there [were] this arrangement, *with no intervening causes*, then the Hall effect [would] occur” (1983, 226, *our italics*). Isolation, purification, control of intervening causes (i.e. control of physical parameters) are noticeable features of an experimental protocol that have a straightforward consequence directly relevant for our philosophical interrogation on serendipity: they tend to limit the number of causal pathways which can influence the response of the object or phenomenon under study experimentally. Unknown causal pathways existing in the real world are thus inoperant (or less operant) in laboratory conditions, thereby limiting the occurrence of unexpected results. Hence our first criterion to evaluate whether a certain system of science favors surprising results: the more the phenomena under study in that system are isolated, purified in highly regimented experimental conditions, the less likely the occurrence of unexpected results is.

Moreover, isolation, purification of phenomena often go hand in hand with another noticeable feature of laboratory sciences, described by Hacking as follows: “as a laboratory science matures, it develops a body of types of theory and types of apparatus and types of analysis that are mutually adjusted to each other” (1992, 30). In particular, a given theoretical framework determines the type of questions that can be probed experimentally, guides the design of apparatus and defines the type of data produced. Consequently, “data uninterpretable by theories are not generated” (Hacking 1992, 55). This process of mutual constraints is well illustrated for instance by recent experimental inquiries in particle physics, such as the quest for the Higgs Boson. Its existence was postulated in the frame of the Standard Model of theoretical physics (Higgs, 1964) and complex experimental apparatus have been developed with the explicit goal of “discovering” it (LEP, 2003). The “discovery” occurred in 2012 (ATLAS, 2012) but the high degree of tailoring of the apparatus to the

theory postulating the particle can be considered as imposing some kind of a priori structure on the phenomenon, so that particles such as the Higgs boson are not so much “discovered” than “manufactured” (Falkenburg, 2007, 53). In any case, the “discovery” of the Higgs boson was hardly a surprise and illustrates Hacking’s more general contention about experimental inquiries typical of contemporary laboratory sciences as opposed to real-world experiments: “[their] results are more often *expected* than *surprising*” (1992, 37, *our italics*).

3.2 Theoretical unifying ambition

Another relevant characteristic of an experimental inquiry is the degree of generality of its theoretical framework. Scientists working within a theoretical framework with a large unifying scope will be reluctant to “leave” it and search for an alternative one when facing an unexpected result, and for good epistemological reasons: there is (obviously) a high epistemic cost of abandoning a theoretical framework that provides explanations for a large set of phenomena. The right move is rather to try to accommodate the surprising result by adopting, if necessary, *ad hoc* hypothesis or tinkering with some ingredients of the existing theoretical framework, so that the result loses its “exteriority” and ends up being integrated. And because of this well-known “plasticity” and integrative power of well-established theoretical frameworks with a large unifying scope², when a (at first sight) surprising result occurs, it rarely leads to the opening up of a new line of inquiry in search of an alternative explanatory framework, but rather gets integrated within the existing one, thereby losing its unexpectedness.

There is another reason why a high degree of theoretical generality does not favor the occurrence of unexpected results, which is linked to our previous remarks on the process of

² Classical references on these ideas of plasticity or integrative power are of course Kuhn’s description (1962) of scientists being busy working on resolving anomalies in normal science and Lakatos’ concept of “protective belt” of a research program (1978).

mutual adjustment between theoretical ingredients, apparatus and data. By constraining the type of experimental procedures developed and the type of data generated, a theoretical framework with a large unifying scope tends to *homogenize* the experimental works conducted to probe the various phenomena that it accounts for. And since a diversity of experimental approaches increases the possible sources of emergence of surprising facts, we can conclude that by reducing this diversity, theoretical generality makes the occurrence of unexpected facts less likely to happen.

The case of the etiology of cancer provides interesting illustrations of these two unexpectedness-diminishing effects of theoretical generality. The classical theory of cancer, the Somatic Mutations Theory (SMT), has been challenged for fifteen years or so by a new theoretical approach, the Tissue Organization Field Theory (TOFT) (Sonnenschein and Soto, 2000). First developed in the 1970's, the SMT rapidly became the dominant research theoretical framework on carcinogenesis (Mukherjee, 2010). This hegemony led to a high degree of homogenization of the experimental inquiries: the experimental procedures were all dedicated to the very standardized search for genetic mutations, in the context of molecular biology. Moreover, many, if not all surprising observations were made compatible with SMT by using *ad hoc* hypothesis (Soto, 2011). For instance, it was observed that various types of cancer were exhibiting large-scale disorganization of the genome. This observation was unexpected to the extent that it could not fit with SMT's fundamental postulate of punctual mutations. To integrate it in the frame of SMT, the existence of an original genetic instability of the cancer cells was then postulated (Rajagopalan, 2003).

4 Use-inspired science, pure science, and unexpected facts

In light of the criteria that we proposed above, how does pure, disinterested science score compared to applied science when it comes to favoring the occurrence of unexpected facts? A

helpful starting point is provided by Martin Carrier's insightful characterization of applied science:

"Three methodological features can be observed whose combined or marked appearance tends to be characteristic of applied science: local models rather than unified theories, contextualized causal relations rather than causal mechanisms, real-experiments rather than laboratory experiment conducted for answering theoretical questions" (2004, 4).

4.1 Local models

Let us start with the contrast between local models and unified theories. Whereas pure science often aims at providing comprehensive and unifying theoretical frameworks (think of the Standard Model in particle physics or the Big Bang model in cosmology), use-inspired research is characterized by the coexistence of numerous local models, each determining the development of specific experimental procedures. An extreme case of this locality are for instance the design-rules used in the industry, which are built as laws guiding action (Wilholt, 2006). They are experimentally confirmed rules providing relations among different relevant parameters to manufacture industrial products. These rules are extremely specific: they apply to a very few number of situations and each of them determines a singular experimental practice. The use of local models is also widespread in the biomedical sciences, a typically use-inspired field of research. We will again draw on oncology to illustrate our point. Consider for instance the case of the development of radiotherapy protocols in the first half of the XXth century. The aim was to intervene on cancer to cure it, without any general model describing the mechanism of carcinogenesis. This program promoted the development of a variety of exploratory approaches using X-rays against cancer (Pinell, 1992). As there were

no standardized protocols, many experimental procedures were tested, changing the density of X-rays received, the distance of emission, the frequency of the radiotherapy sessions. In order to improve the efficiency of the therapeutic methods, scientists tried to build various local models describing the action of X-rays on cancer, corresponding to the variety of experimental procedures implemented. Grubbe (1949) formulated a model based on the inflammatory reaction to explain the effects of radiotherapy on cancer: the inflammation of the surrounding tissue beyond the effects of X-rays is responsible for the decrease of tumoral mass. This model is applicable to his specific use of X-rays: he applied very high doses, necessary to generate an inflammatory response. In parallel, Tribondeau and Bergonié, using more moderate doses, developed a model based on the proliferation of the cells in tumoral context, which led to the "Bergonié law": X-rays have a higher impact on proliferating cells (Tribondeau, 1959).

What lessons can be drawn from this first contrast between local models and unified theories? The answer is rather straightforward, given the link spelled out in the previous section between the level of generality of theoretical models and the occurrence of unexpected facts (our second criterion): by promoting the use of a diversity of local models and heterogeneous experimental protocols, applied science favors the occurrence of unexpected facts, whereas the penchant of pure science for comprehensive unifying theoretical frameworks, hence homogenized experimental protocols, does not.

4.2 Causal incompleteness

Let us compare now pure science and applied science in light of our first criterion based on the degree of isolation and purification of the phenomena under study. A directly relevant feature of applied science is the use of what Carrier calls "contextualized causal relations" rather than full causal chains. Use-inspired science typically aims at directly intervening on a

process or phenomenon often disposing only of a partial knowledge of the causal chains involved and without being able to isolate it from various causal influences exerted by the rest of the physical world. A direct consequence of this feature of applied science is the low degree of control of its experimental protocols. By contrast, since pure science aims primarily at answering fundamental theoretical questions, it designs highly regimented experimental procedures that isolate and purify phenomena in order to be able to get empirical answers about the specific fundamental processes questioned in the theoretical investigation³. Moreover, building highly regimented experimental procedures requires knowledge of full causal chains in order to be able to better control the response of the system under study. The outcome of the application of our criterion is then again straightforward: compared with pure science, applied science favors the occurrence of unexpected facts to the extent that its experimental procedures are less controlled and based only on partial knowledge of the causal influences exerted on the phenomenon under study.

The etiology of cancer provides again interesting illustrations of our claim. Indeed, many current cancer therapies built in the frame of use-inspired research are based on contextualized causal relations. Typically, if a cellular agent is found to be massively expressed in cancer cells, drugs are designed to inhibit it, even if the whole causal chain determining its action is not known. For instance, a large amount of proteins promoting angiogenesis (the growth of blood vessels), notably VEGF (Vascular Endothelial Growth Factor), was found in tumoral cells, leading to the design of anti-VEGF molecules (Sitohy,

³ Carrier sums up this contrast as follows: "Empirical tests often proceed better by focusing on the pure cases, the idealized ones, because such cases typically yield a more direct access to the processes considered fundamental by the theory at hand. But applied science is denied the privilege of epistemic research to select its problems according to their tractability (...). Practical challenges typically involve a more intricate intertwinement of factors and are thus harder to put under control". (2004a, 4) In the life sciences, this focus on "pure cases" means using "model organisms" or a limited number of well spread cell lines (e.g. the HeLa cells or the *Saccharomyces Cerevisiae* yeast) to elucidate fundamental biological mechanisms. And the use of such standardized objects tends to homogenize the experimental protocols.

2012). These molecules are used without considering the complete causal chain in which the VEGF is embedded. Only their known action on angiogenesis is considered. The clinical tests have led to unexpected observations: the use of an anti-VEGF molecule (Avastin) can stimulate tumor growth (Lieu *et al.*, 2013)⁴. This example shows that the use of contextualized causal relations promotes the appearance of surprising facts by allowing unknown mechanisms to intervene in the experimental procedure.

5. Concluding discussion

Our previous analysis has established that several features of pure, disinterested science make it less hospitable than use-inspired science to the occurrence of unexpected facts. For all that, it does not follow that proponents of freedom of science cannot appeal anymore to unpredictability in the sense of serendipity to make their case. For the issue of which conditions favor the occurrence of unexpected facts is only half of the story. The other half is the possibility to actually follow up on these occurrences and open new lines of inquiry. And this other half raises different issues. What are the institutional, social structures of science that make it easier for scientists to re-orient their research when needed? To what extent an initial orientation of a scientific investigation by “external” practical needs is less compatible with the opening of new lines of inquiry than an initial orientation by epistemic considerations internal to the dynamics of a scientific field? When appealing to the serendipity argument,

⁴ Interestingly, this observation led to new use-inspired research programs, aiming at identifying the molecular causal pathways giving rise to this tumoral resistance phenomenon. It has notably strongly oriented the research toward the precise understanding of the VEGF pathways (Moens, 2014). For instance, the study of the mechanisms of expression in cancer cells of various kinds of VEGF agents is becoming an important program of research (Li, 2014) and these works allow to build new fundamental knowledge about the action of the VEGF proteins.

proponents of free, disinterested science not only presuppose that it is the best system of science to generate unexpected facts to start with – a contention that we have challenged in this paper – but also that it actually gives more freedom to scientists to follow up on unexpected results. In other words, the issue of scientists' given possibility to change the direction of their research when needed is somewhat mixed, confused with the normative issue of what the aims of science should be (in short, increase knowledge following considerations internal to science *vs.* answer external practical needs). But it seems to us that the two issues should be kept separate. After all, one can very well conceive a system of science whose aims are primarily to answer society needs but which nevertheless leaves scientists free to choose the lines of inquiry that seem *to them* the most promising ways of fulfilling these needs (which includes changing research directions if needed). Otherwise put, one can very well conceive a use-inspired science which is not a *programmed* science in which scientists are asked to plan every step of their inquiry in order to achieve a given aim. And note that a pure, disinterested science may be as much programmed as a use-inspired science: the fact that scientists are left free to choose the aims of their research does not protect them from having to plan every step to reach these aims. In any case, our purport in this paper was not to attack pure, disinterested science. There are, no doubt, many good reasons to defend it, but the widespread, traditional one grounded on the unpredictability of scientific inquiry is certainly not the most epistemologically cogent and solid one.

REFERENCES

ATLAS Collaboration. 2012. "Observation of a new particle in the the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Phy. Lett. B* 716(1) :1-29

- Becquerel, Henri. 1896. "Sur les radiations émises par phosphorescence". *Comptes-rendus de l'Académie des sciences*.
- Brooks, Harvey. 1994. "The relationship between science and technology". *Research Policy* 23(5):477-86
- Bush, Vannevar. 1945. *Science, The Endless Frontier. A Report to the President by Vannevar Bush, Director of the Office of Scientific Research and Development*. Washington D. C.: National Science Foundation.
- Carrier, Martin. 2004. "Knowledge and Control: On the Bearing of Epistemic Values in Applied Science". In *Values and Objectivity in Science*, ed. P. Machamer and G. Wolters, 275-293. Pittsburgh, PA: University of Pittsburgh Press.
- — —. 2004a. "Knowledge gain and practical use: Models in pure and applied research". In *Laws and Models in Science*, ed. D.Gillies, 1:17. London: King's College Publications
- Edgerton, David. 2004. "The Linear Model Did not Exist. Reflections on the History and Historiography of Science and Research in Industry in Twentieth Century". In *Science-Industry Nexus: History, Policy Implications*, ed. Karl Grandin and Nina Wormbs, 31-57. New-York: Watson.
- Falkenburg, Brigitte. 2007. *Particles Metaphysics. A critical Account of Subatomic Reality*. Springer.
- Fert, Albert. 2007. Interview published in *Le Monde*, October, 25, 2007.
- Flemings, Alexander. 1929. "On the antibacterial action of cultures of a penicillium with special reference to their use in the isolation of b. influenza". *J. Exp.Path.* 10:226-36.
- Grubbe, Emil. 1949. *X-Ray Treatment: Its Origins, Birth, and Early History*. St.Paul and Minneapolis, MN: Bruce Publishing Company.
- Hacking, Ian. 1983. *Representing and Intervening*. Cambridge University Press.
- — —. 1992. "The Self-Vindication of the Laboratory Sciences". In *Science as*

- practice and culture*, ed. A. Pickering, 29-64. The University of Chicago Press.
- Higgs, Peter W. 1964. "Broken Symmetries and The Masses of the Gauge Bosons".
Phys.Rev. Lett 13:508.
- Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. The University of Chicago Press.
- Lakatos, Imre. 1978. *The methodology of scientific research programs*. Cambridge: Cambridge University Press.
- LEP Collaboration. 2003. "Search for the Standard Model Higgs Boson at LEP". *Physics Letters B* 565:61-75
- Leydesdorff, Loet and Etzkowitz Henry. 1997. *Universities and the Global Knowledge Economy: A Triple Helix of University-Industry-Government Relation*. London, Cassel Academic.
- Li, Dong et al. 2014. "Tumor resistance to anti-VEGF therapy through up-regulation of VEGF-C expression". *Cancer Lett* 346:45-52.
- Lieu, Christopher H. et al. (2013). The association of alternate vegf ligands with resistance to anti-vegf therapy in metastatic colorectal cancer. *PLoS One* 8(10):e77117.
- Moens, Stijn et al. 2014. "The multifaceted activity of VEGF in angiogenesis - Implications for therapy responses". *Cytokine Growth Factor Rev* 25:473-82.
- Mukherjee, Siddhartha. 2010. *The Emperor of All Maladies. A Biography of Cancer*. * Scribner.
- Pinell, Patrice. 1992. *Naissance d'un fleau. Histoire de la lutte contre le cancer en France (1890-1940)*. Métailié.
- Polanyi, Michael. 1962. "The Republic of Science: Its Political and Economic Theory".
Minerva 1: 54-73.
- Rajagopalan, Harith, Nowak Martin A, Vogelstein Bert and Langauer Christoph. 2003. "The

- significance of unstable chromosomes in colorectal cancer". *Nat Rev Cancer* 3(9):695-701.
- Rosenberg, Nathan. 1992. "Science and Technology in the Twentieth Century". In *Technology and Enterprise in Historical Perspective*. Oxford, Clarendon Press.
- Rosenberg, Barnett *et al.* 1967. "The inhibition of growth or cell division in escherichia coli by different ionic species of platinum(iv) complexes". *J. Biol.Chem* 242(6):1347-5.
- Sitohy, Basel. 2012. Anti-vegfr therapy for cancer: reassessing the strategies. *Cancer Res* 8:1909-14.
- Sonnenschein, Carlos and Soto A- M. 2000. "Somatic mutation theory of carcinogenesis: why it should be dropped and replaced". *Mol. Carcinog.* 29(4):205-211.
- Soto, Ana M. and Sonnenschein C. 2011. "The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory". *Bioessays* 33(5):332-340.
- Stokes, Donald E. 1997. *Pasteur's Quadrant. Basic Science and Technological Innovation*. The Brookings Institution.
- Tribondeau, Jean B. 1959. "Interpretation of some results of radiotherapy and an attempt at determining a logical technique of treatment". *Radiation Research*, 11(4):587-588.
- Wilholt, Torsten. 2006. "Design rules: Industrial research and epistemic merit". *Philosophy of Science* 73(1):66-89.
- Wilholt, Torsten and Glimell H. 2011. "Conditions of Science: The Three-Way Tension of Freedom, Accountability and Utility". In *Science in the Context of Application*, eds. M.Carrier and A.Nordmann, 351-70. Boston Studies in the Philosophy of Science.

(Accepted for publication in *Philosophy of Science*,
subject to revision after presentation at 2016 PSA meeting)

Using Democratic Values in Science: an Objection and (Partial) Response¹

*S. Andrew Schroeder (aschroeder@cmc.edu),
Claremont McKenna College*

draft of June 2016

Abstract

Many philosophers of science have argued that social and ethical values have a significant role to play in core parts of the scientific process. A question that naturally arises is: when such value choices need to be made, *which* or *whose* values should be used? A common answer to this question turns to political values — i.e. the values of the public or its representatives. In this paper, I argue that this imposes a morally significant burden on certain scientists, effectively requiring them to advocate for policy positions they strongly disagree with. I conclude by discussing under what conditions this burden might be justified.

1. Values in Science and the Political View

By now, most philosophers of science probably agree that there is an important place for so-called contextual (i.e. personal, ethical, political) values in core parts of the scientific process, especially in areas where science is connected to policy-making. Values may appropriately play a role in evaluating evidence (Douglas 2009), choosing scientific models (Elliott 2011), structuring quantitative measures (Reiss 2013, ch. 8; Stiglitz, Sen, and Fitoussi 2010; Hausman

¹ For comments on earlier drafts of this paper, I thank Alex Rajczi and the students in a seminar on science and values at Claremont McKenna College. For discussions on related topics, I thank Gil Hersch, Daniel Steel, and Branwen Williams. This work was supported in part by a research grant from the Claremont McKenna College Center for Innovation and Entrepreneurship.

2015), and/or in preparing information for presentation to non-experts (Elliott 2006; Hardwig 1994; Resnik 2001; Schroeder 2016). The natural follow-up question has received less sustained attention: when scientists should make use of values, *which* (or *whose*) values should they use?²

In some cases, philosophers of science criticize a value choice on substantive ethical grounds (e.g. Shrader-Frechette 2008; Hoffmann and Stempsey 2008). This suggests that the values to be used are the objectively correct ones. A second common view gives scientists latitude to choose whatever (reasonable) values they prefer or think best, usually supplemented by a requirement of transparency. This is suggested by many existing codes of scientific ethics, which impose few constraints on scientists in making such choices.³ Finally, a third view says that scientists ought to use the appropriate political values — that is, the values held or endorsed by the public or its representatives — at least when those values are informed and substantively reasonable.⁴ The most straightforward argument for this view grounds it in considerations of democracy or political legitimacy. If certain value choices are going to ultimately influence policy, then the public or its representatives have a right to make those choices (Douglas 2005; Intemann 2015; *cf.* Steele 2012; Kitcher 2001).

There are, of course further possibilities, and these views can be combined in more complex ways (e.g. requiring scientists to use political values in some domains, while permitting them to use their personal values in others). But if, for simplicity, we stick to these three primary

² In some cases, the justification for incorporating values into the scientific process dictates an answer. Feminist critiques of historically androcentric fields, for example, suggest that non-androcentric values are needed as a corrective. I set aside such cases in this paper.

³ Mara Walli, Matthew Wong, and I discuss this at length in a work-in-progress.

⁴ I set aside, then, cases where the values, say, of a policy-maker are unreasonable, in the sense that they lie outside the range of values that ought to be tolerated in a liberal society. In such cases, an advocate of the political view may permit or require scientists to reject those unreasonable values. (See e.g. Resnik 2001.) Also, in this paper I will set aside the important question of what the political view ought to say when the values of the public diverge from the values of policy-makers. The answer to this question, I think, will depend on one's theory of political representation.

options, I think the third, which I will call the *political view*, is the most attractive. More precisely, I think that in most cases where values are called for in core parts of the scientific process, scientists should privilege political values.⁵ The most obvious concern with this view, and one that has received much attention from its advocates, is that it doesn't seem practical. It isn't feasible to ask citizens or policy-makers to weigh in at every point in the scientific process where values are required, and even if we could, non-experts often will not have the scientific background to fully understand the options before them. Substitutes for actual participation on the part of policy-makers or the public, such as asking scientists to predict what the public would choose or to determine what values policy-makers would hold upon reflection, seem to place unreasonable epistemic demands on scientists.

Douglas (2005), Intemann (2015), Guston (2004), and others have argued that these problems aren't insurmountable, by suggesting specific ways that the concerns of policy-makers and the public can be brought into the scientific process. And Kevin Elliott (2006; 2011) has suggested a more general way we might make progress. The political view goes hand-in-hand with a view of the relationship between science and policy that is widely-held: that the role of a scientist is to promote informed decision-making by policy-makers.⁶ Bioethicists have extensively discussed how health care professionals can promote informed decision-making on the part of patients and research subjects. Theoretical and empirical research has led to a range of suggestions for how physicians can promote informed decision-making, even in cases where a patient's values may be uncertain, different research subjects may hold different values, and so

⁵ This, of course, is proposed as a principle of professional ethics - not e.g. a legal requirement.

⁶ See also Resnik (2001), Martin and Schinzinger (2010), and Schroeder (2016) for theoretical defenses of this idea, which is consonant with the mission statements of many scientific organizations and associations.

forth. Elliott's hope is that many of these suggestions can be adapted to the scientific case, or at least a parallel research program could be carried out, informed by the work of bioethicists.⁷

It is, of course, far from established that these proposals will work, but the range of options on the table strikes me as cause for optimism. And even if these solutions don't work in all cases, there is still bite to the political view, since it could still tell scientists to use political values *when they can determine those values*. Accordingly, in this paper I would like to describe a different and I think deeper concern with the political view, one which has been conspicuously absent from the literature thus far. In requiring scientists to guide certain aspects of their work by political values, we will sometimes in effect ask that they support political causes they may personally oppose and bar them from fully advocating for their preferred policy measures. We are, then, depriving scientists of important political rights possessed by the general public. In the remainder of this paper, I will spell out this objection more fully and explain why I think it has significant moral force. In the end, I will suggest that although there is reason to think that the objection doesn't ultimately undermine the political view, it nevertheless constitutes a significant cost that accompanies that view, which its proponents need to acknowledge.

2. Two Cases Where the Political View Seems Troublesome

The literature on values in science is vast and diverse, and so it will be useful to have some particular examples in mind. First, consider Douglas's (2000; 2009) argument that scientists should or must appeal to value judgments when resolving certain uncertainties that arise during the scientific process. Scientists conducting research into the potential carcinogen dioxin, for example, were faced with liver samples which had tumors that could not clearly be

⁷ See also Schroeder (2016) for how this might go.

categorized as malignant or benign. In resolving such borderline or ambiguous cases, Douglas argues that scientists should appeal to contextual values, when the constitutive norms of science don't dictate any resolution. In this case, health-protective values would lead scientists to classify borderline samples as malignant; while concerns about overregulation would lead scientists to classify those same samples as benign (Douglas 2000).

Second, consider the many choices that scientists have to make when preparing their results for presentation. How should uncertainty be characterized? (Should 90% or 95% confidence intervals be used?) Which study results should be highlighted? (Which drug side effects should be discussed at length, and which included as part of a long list?) How should statistics be summarized? (As means or medians? Should results be broken down by gender, or presented only in aggregate?) In making choices like these, scientists frequently must appeal to values — to decide, for example, which pieces of information are important and which are not.⁸

It is, I presume, fairly uncontroversial that these value choices — how to resolve uncertainties in the research process and how to present results — can influence policy in foreseeable ways. Douglas, for example, argues that this is the case in the dioxin studies. Classifying borderline samples as malignant will make dioxin appear to be a more potent carcinogen, likely leading policy-makers to regulate it more stringently (2000, 571). Keohane, Lane, and Oppenheimer (2014) show how a presentation choice made by the Intergovernmental Panel on Climate Change led to poor policy outcomes, which likely could have been avoided by presenting information differently. More generally, we know from a wealth of studies in psychology and behavioral economics that the way information is presented to someone can strongly influence her subsequent choices (Thaler and Sunstein 2008), and there have been

⁸ For discussions, see Elliott (2006), Hardwig (1994), Keohane, Lane, and Oppenheimer (2014), Resnik (2001), and Schroeder (2016).

several influential commentaries calling for scientists to more carefully “frame” their results (Nisbet and Mooney 2007; Lakoff 2010). So it seems straightforward that the value choices made by scientists can predictably affect policy.

If these value choices can influence policy, then in directing scientists to make them in accordance with political values — as opposed to the scientists’ personal values — we are asking scientists to characterize policy-relevant material in a way that may promote an outcome they strongly disfavor. For example, suppose the scientists in Douglas’s dioxin study value public health much more than they value keeping industry free from overregulation, but the public and its elected representatives have the opposite view. Further, suppose both views are substantively reasonable, in that they are within the range of policies eligible for adoption through democratic processes. In this case, the political view would tell the scientists to categorize borderline samples as benign, since that would better cohere with the public’s values. This could make dioxin appear to have minimal carcinogenic effects, predictably leading to less regulation than would have occurred had the scientists classified borderline samples according to their own, health-protective values. Similarly, suppose an environmental economist conducting an impact study of a proposed construction project is herself deeply committed to the preservation of natural spaces. Nevertheless, if the public is strongly committed to economic development, the political view would require her to put front-and-center a detailed breakdown of the economic consequences of construction, while describing the ecological costs more briefly or in a less prominent place — likely frustrating her desire for preservation.

Notice that the concern here is not simply that scientists are being asked to provide information that will lead to an outcome they disfavor. I take it that any reasonable approach to scientific ethics will require that scientists communicate honestly, even in cases where that

promises to yield policies they don't like. Similarly, I presume that scientists must also be forbidden from presenting information in ways that, though technically accurate, are nevertheless misleading. The problem here is that Douglas's scientists are being asked to characterize results in one way (as benign) that could, *with equal scientific validity*, have been characterized differently (as malignant). And our environmental economist is being asked to present her results in one way (highlighting economic benefits), when an alternate presentation (one highlighting ecological costs) would be equally honest, accurate, objective, transparent, clear, and so forth. In each case, then, we have a collection of underlying data which can be described or characterized in different ways, neither of which appears to be more scientifically valid than the other. The political view insists that scientists choose the description grounded in values they don't accept and which seems likely to promote policy outcomes they disfavor. In this respect, the political view requires scientists to in effect advocate for, or at least tilt the playing field towards, political views they disagree with.⁹

3. Elliott and The Principle of Helpfulness

This seems clearly to be a significant imposition on scientists and thus a cost of the political view. It is therefore surprising that, so far as I can tell, philosophers who have argued for the political view have not commented on it. This is most striking in Elliott's work. Elliott, recall, argues that scientists should aim to promote informed decision-making among policy-makers, in something like the way physicians should aim to promote informed decision-making among patients. Standard accounts in bioethics say that it is the patient's values that carry the

⁹ Can't we let the scientists advocate for their preferred positions in other ways? We could let scientists present their preferred interpretation separately. But if the political view is to have bite, presumably these alternate results will have to be clearly designated so and offered in a less prominent place (e.g. in an appendix or online supplement). And we should of course permit scientists to advocate for their views outside of their scientific papers/reports. But it seems likely that these (private) statements will carry much less policy weight than their scientific ones.

day: in normal cases, the physician's job is to help a patient make decisions that cohere with her own values. If the scientific cases is analogous, then the scientist's job is to help policy-makers make decisions that cohere with their (or the public's) values. This, in turn, suggests that scientists should use political values when resolving uncertainties, presenting results, and so forth. In other words, Elliott's proposal seems to imply the political view.¹⁰

The main defense Elliott offers for this view, however, relies on Scanlon's "Principle of Helpfulness":

Suppose I learn, in the course of conversation with a person, that I have a piece of information that would be of great help to her because it would save her a great deal of time and effort in pursuing her life's project. It would surely be wrong of me to fail (simply out of indifference) to give her this information when there is no compelling reason not to do so.¹¹

Elliott sums up the idea this way: "[I]n situations where one can significantly help another individual by engaging in an action that requires little sacrifice, it is morally unacceptable not to help" (2011, 139). If the political view, however, requires characterizing data or presenting information in ways that promote policy choices a scientist strongly opposes, then this Principle doesn't apply. When the pro-health scientist is required to classify ambiguous samples as benign, that does involve a sacrifice. A refusal to do so — which would hinder the pro-industry policy-maker's ability to make an informed regulatory decision — would not be done "simply out of indifference". It would be done out of the scientist's desire to protect public health.

¹⁰ In some work, Elliott appears to suggest that transparency about values may be enough (Elliott and Resnik 2014). That is, he doesn't seem to place (many) constraints on scientists' value choices, so long as they are open about those choices. If that is Elliott's view — and it is not clear to me that it is — it strikes me as in tension with his insistence that scientists promote informed decision-making. Surely I can better help you make a decision that coheres with your values by working from your values, rather than by working from my own values (even if I am open about what I am doing). Further, even if scientists are open about their value choices, policy-makers frequently won't have the technical expertise to be able to reinterpret a scientific study, replacing one set of values (the scientist's) with another (their own). (If values could so easily be swapped out by non-specialists, then much of the debate about values would be unimportant. Transparency is all we would require.)

¹¹ Scanlon (1996, 224), quoted in Elliott (2011, 139).

(Similar things, obviously, can be said about the environmental economist asked to highlight the economic aspects of a proposed construction project.)

Scanlon's Principle of Helpfulness is a quite weak one, applying only in cases where the agent in question can put forward no significant burden of compliance. That Elliott uses it to justify his informed decision-making framework, and implicitly the political view, suggests that he thinks that such a view doesn't impose significant burdens on scientists. But if what I've said has been correct, that is wrong. Even if the political view is justified — and, as I've said, I think it is — we need to recognize that it asks a lot of scientists in cases where their values diverge from those of the relevant political body.

4. Physicians vs. Scientists

This, however, brings up an interesting question. If Elliott is right that the scientific case is analogous to the biomedical case, then shouldn't informed consent requirements in medicine be treated as similarly burdensome? Few bioethicists, though, would have sympathy for a physician who claimed that seeking informed consent constituted a significant ethical burden. (They may have sympathy for the claim that seeking informed consent is burdensome in more mundane ways — e.g. too time-consuming — but those complaints seem very unlike the scientists'.) I think that there is an important difference between the cases, which will help us to more clearly understand why the scientist is often burdened in a way that carries moral weight, while the physician normally is not.

We can see this by constructing a case which seems to put a physician in a position like the scientist's. Consider Jane, a doctor who strongly believes that the end of life for terminal patients is greatly enhanced by effective pain management, even if doing so shortens the

patient's life or impairs his consciousness. For this reason, Jane has chosen palliative care as her specialty, making it her life's work to help dying patients avoid unnecessary pain. One of her patients, John, has continually insisted that he wants to remain as lucid as possible, even if that means agony. As he lies here, in agony, Jane suspects that if she framed the information properly — highlighting a medication's ability to relieve pain, while downplaying its cognitive effects — she might be able to get John to accept it. And accepting the medication, Jane strongly believes, would be much better for John. Nevertheless, standard interpretations of informed consent forbid her from doing so. Knowing that John is especially concerned about lucidity, she is ethically bound to highlight that information when informing him of his options. Unsurprisingly, John declines the pain medication and experiences what Jane regards as an awful death — precisely the kind of thing she went into palliative care to prevent.

Like our pro-health scientist, Jane has been asked to present information in a way that ultimately frustrates her deeply-valued goals. But imagine Jane complains to the ethics board at her hospital, arguing that it is burdensome to ask her to highlight to John the effects of pain medication on lucidity, because doing so would frustrate her deeply-held values. This complaint doesn't strike me as at all compelling. Why? Because Jane's values shouldn't hold any sway over John's medical choices. John has the right to reject pain medication, whatever Jane (or just about anyone else) thinks about it. Put another way, John has no obligation to take Jane's wishes into consideration, when he makes his decision. His decision is ultimately *his*.

Now, imagine our pro-health scientist complains to her ethics committee, asserting that it is burdensome to ask her to present her data in a pro-industry light, when it could with equal scientific validity be presented in a pro-health light, because doing so would frustrate her deeply-held concern for public health. Or imagine the environmental economist complaining about

having to foreground the economic benefits of the proposed construction project, since doing so will make it more likely that the project is approved and another natural space will be bulldozed. If we assume that the scientists are citizens of the society in question, then their situation is different from Jane's. As citizens in a democracy, their views should hold some sway over their government's policy choices. A government does have an obligation to take its citizens' views into consideration when making policy decisions. And when the government ultimately acts, it does so on the scientists' behalf. The decision is, in part, the scientists'.

The scientists, then, are stakeholders and even part-decision-makers in the associated policy-decisions, in a way that Jane is not a stakeholder in John's decision. This is true even if Jane cares more about John's decision than our scientists care about the policy decisions. We can see, then, that the political view isn't burdensome simply because it directs scientists to promote or advocate for outcomes they disfavor. It is burdensome because it sometimes directs scientists to promote or advocate for disfavored views, on matters that they have a right to speak on, to a body that purports to act on their behalf. This is what gives their burden its moral significance.¹²

5. Justifying the Burdens of the Political View

Some scientists have recognized the burdens that even neutrality — let alone the political view — would impose on them.

Conservation biology is inescapably normative. Advocacy for the preservation of biodiversity is part of the scientific practice of conservation biology. If the editorial policy of or the publications in [the journal] *Conservation Biology* direct the discipline toward an “objective, value-free” approach, then they do not educate and transform society... To pretend that the acquisition of “positive knowledge” alone will avert mass extinctions is misguided... Without openly acknowledging such a perspective,

¹² What about cases where the scientists are not citizens of the society in question? In some cases, we can still make out a stakeholder claim. (When it comes to climate change, for example, we are all stakeholders in U.S. climate policy.) But such cases raise complications which I unfortunately can't discuss in a short paper like this one.

conservation could become merely a subdiscipline of biology, intellectually and functionally sterile and incapable of averting an anthropogenic mass extinction. (Barry and Oelschlaeger 1996)¹³

Most conservation biologists enter that field because of a strong commitment to the value of biodiversity and the preservation of nature (Marris 2006). Similar things are surely true of other scientific disciplines. (My experience has been that public health researchers and economists studying inequality disproportionately share certain political values.) To the extent that these values diverge from the values of the public and its representatives, the political view would require these scientists to continually characterize their results in ways structured by a value system they find unacceptable. (In this respect, things would be quite different for, say, climate scientists. Although their work is controversial, it nevertheless is founded on values that are widely shared. The potentially catastrophic consequences of climate change are ones that virtually everyone cares about. Climate change deniers typically object to the *empirical* claims made by climate scientists - not to the basic values they hold.)

Is it fair, then, to tell a conservation biologist, who perhaps entered the field because of her love for natural spaces and has spent the bulk of her life collecting information that she hopes can be used to preserve them, that she is nevertheless ethically bound to resolve uncertainties in her research in ways favorable to economic growth, or to present her results in ways that highlight the economic value (as opposed to, say, the private or aesthetic value) of undeveloped land? I don't have a full answer to this question — such an answer would require more empirical information, as well as a fuller discussion of political philosophy — but I think we can see how the argument would go. There are a range of situations in which we impose significant

¹³ This article was followed by a collection of commentaries, most of which generally supported the authors' views. Similar proposals seem to crop up frequently among conservation biologists, and are generally endorsed by those in the field (Marris 2006).

restrictions on speech and advocacy for people in important social positions. The Code of Conduct for U.S. judges, for example, bars judges from publicly endorsing candidates for political office and from making speeches for political organizations.¹⁴ Uniformed U.S. military personnel are not permitted to participate in political fundraising, speak at political events, or display political signs, even on their private vehicles.¹⁵ Other constraints on speech and advocacy seem ethically appropriate for politicians, police officers, lawyers, and others.

So, if there is an important public good served by constraining scientists' advocacy, it doesn't seem in principle problematic to do so. Two arguments along these lines seem promising. First, a distinctly political approach might argue that although imposing this burden on scientists does restrict important political rights of speech and advocacy, it is done in order to expand the political rights of others. By requiring scientists to work from the values of the public, the ability of the public to make informed policy choices and to effectively advocate for their own positions is enhanced. Thus, although the political view constitutes a loss of political freedom to scientists, that loss is more than balanced by the gain in political freedom to the public as a whole. (A view like this seems generally consistent with an approach to democracy like Brettschneider's (2007).)

Second, a straightforwardly consequentialist argument could point out the terrible consequences that threaten to follow if the public and/or policy-makers distrust scientific results. One of the primary arguments that has been put forward in favor of informed-consent approaches in bioethics has been that it promotes trust on the part of patients. Similarly, Elliott's informed decision-making approach — which implies the political view — seems like a promising way to

¹⁴ <http://www.uscourts.gov/judges-judgeships/code-conduct-united-states-judges>

¹⁵ <http://www.dtic.mil/whs/directives/corres/pdf/134410p.pdf>

promote trust in science (Elliott 2011, 133-6; *cf.* Hardwig 1994; Resnik 2001). If, then, the political view proves to be an effective way of promoting public trust in science, which in turn heads off the problems that ensue when policy-makers disregard science, that could justify imposing significant burdens on scientists.

Neither of these defenses, of course, is anywhere near complete. But both do strike me as quite reasonable, and so I don't think the concerns I've discussed in this paper should lead proponents of the political view to give up that position. That said, it is important to note the form that these defenses take. Neither attempts to show that the burden on scientists is not morally significant (as, perhaps, we might be inclined to say about the complaint of the palliative care physician). Instead, they each point to compensating benefits — not necessarily enjoyed by the scientists in question — which morally outweigh the scientists' burden. This means that the political view, even if it is justified, comes at a real cost to scientists, which is something its proponents need to acknowledge.

References

- Barry, Dwight and Max Oelschlaeger. 1996. "A Science for Survival: Values and Conservation Biology," *Conservation Biology* 10: 905-11.
- Brettschneider, Cory. 2007. *Democratic Rights: The Substance of Self-Government*. Princeton University Press.
- Douglas Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67: 559-79.
- Douglas, Heather. 2005. "Inserting the Public Into Science." In *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*, ed. Sabine Maasen and Peter Weingart, 153-169. Springer.
- Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.
- Elliott, Kevin C. 2006. "An ethics of expertise based on informed consent." *Science and Engineering Ethics* 12: 637-61.
- Elliott, Kevin C. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. Oxford University Press.
- Elliott, Kevin C. and David B. Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives* 122: 647-50.
- Guston, David. 2004. "Forget Politicizing Science. Let's Democratize Science!" *Issues in Science and Technology* fall 2004.
- Hardwig, John. 1994. "Toward and Ethics of Expertise." In *Professional Ethics and Social Responsibility*, ed. Wueste, 83-101. Roman and Littlefield.
- Hausman, Daniel. 2015. *Valuing Health: Well-Being, Freedom, and Suffering*. Oxford University Press.
- Hoffman, George and William Stempsey. 2008. "The Hormesis Concept and Risk Assessment: Are There Unique Ethical and Policy Considerations?" *BELLE Newsletter* 14: 11-17.
- Intemann, K. 2015. "Distinguishing between legitimate and illegitimate values in climate modeling." *European Journal for Philosophy of Science* 5: 217-32.
- Keohane, Robert O., Melissa Lane, and Michael Oppenheimer. 2014. "The ethics of scientific communication under uncertainty." *Politics, Philosophy & Economics* 13: 343-368.
- Kitcher, Phillip. 2001. *Science, Truth, and Democracy*. Oxford University Press.
- Lakoff, George. 2010. "Why it Matters How We Frame the Environment." *Environmental Communication* 4: 70-81.
- Marris, Emma. 2006. "Should conservation biologists push policies?" *Nature* 442: 13.
- Martin, Mike and Roland Schinzinger. 2010. *Introduction to Engineering Ethics (2nd ed.)*. McGraw-Hill.
- Nisbet, Matthew and Chris Mooney. 2007. "Framing Science." *Science* 316: 56.
- Reiss, Julian. 2013. *Philosophy of Economics: A Contemporary Introduction*. Routledge.
- Resnik, David. 2001. "Ethical Dilemmas in Communicating Medical Information to the Public." *Health Policy* 55: 129-49.
- Scanlon, Thomas. M. 1996. *What We Owe to Each Other*. Harvard University Press.
- Schroeder, S. Andrew. 2016. "Communicating Scientific Results to Policy-Makers." Paper presented at the American Philosophical Association Conference (Pacific Division). Available at <<http://apa-pacific.org/framed/download.php?file=200.pdf>>.
- Shrader-Frechette, Kristin. 1994. *Ethics of Scientific Research*. Rowman and Littlefield.
- Shrader-Frechette, Kristin. 2008. "Ideological Toxicology: Invalid Logic, Science, Ethics About Low-Dose Pollution." *BELLE Newsletter* 14: 39-47.
- Steele, Katie. 2012. "The Scientist qua Policy Advisor Makes Value Judgments." *Philosophy of Science* 79: 893-904.
- Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. 2010. *Mis-measuring Our Lives: Why GDP Doesn't Add Up*. The New Press.
- Thaler, Richard and Cass Sunstein. 2008. *Nudge*. Yale University Press.

Two Roads Diverge in a Wood: Indifference to the Difference Between ‘Diversity’ and ‘Heterogeneity’ Should Be Resisted on Epistemic and Moral Grounds

Anat Kolumbus*, Ayelet Shavit* and Aaron M. Ellison

”
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference

from *The Road Not Taken*, by Robert Frost (1916)

Abstract:

We argue that a conceptual tension exists between “diversity” and “heterogeneity” and that glossing over their differences has practical, moral, and epistemic costs. We examine how these terms are used in ecology and the social sciences; articulate a deeper linguistic intuition; and test it with the *Corpus of Contemporary American English (COCA)*. The results reveal that ‘diversity’ and ‘heterogeneity’ have conflicting rather than interchangeable meanings: heterogeneity implies a *collective* entity that *interactively integrates* different entities, whereas diversity implies *divergence*, not integration. Consequently, striving for diversity alone may increase social injustice and reduce epistemic outcomes of academic institutions and governance structures.

* Equal main contributors.

Key words: collectivity, diversity, ecology, heterogeneity, injustice, institutional diversity.

Acknowledgments: We deeply thank the many different scholars, from very different disciplines, whose work and ideas helped us develop the ideas in this paper. In particular we want to mention Tal Israeli, Tamar Sovran, Nadav Sabar, Daryl G. Smith and Elihu Gerson. They all responded to a single email from an anonymous B.A. student with the same rigor, enthusiasm and respect as to an established full professor, and thus demonstrated the true spirit of academic inclusiveness this paper seeks to explicate. AS's work is supported by Tel Hai College and the ISF (Israeli Science Foundation) grant 960/12 and AME's work on diversity, heterogeneity, and inclusivity in science is supported by the Harvard Forest, and by grant DBI 14-59519 from the US National Science Foundation..

1. Introduction: Diversity in the Ecological and Social Sciences

The concepts of diversity and heterogeneity are two basic types of dissimilarity that are implicitly and commonly assumed to hold interchangeable meanings by scholars and laymen alike. However, when we examined their actual usage, a surprising conceptual discrepancy – in fact a tension – emerged. In this article we call attention to this tension between 'diversity' and 'heterogeneity'¹ and we argue that there are non-trivial epistemic, moral, and practical costs to science and society when this difference is glossed over.

Our critical examination is part of a large body of literature on the benefits of diversity for science and society. There exist strong epistemic (Shrader-Frechette 2002; Longino 2002; Solomon 2006b) and moral (Haraway 1979; Fricker 2007; Douglas 2009, 2015) arguments for diversity in institutions, governance structures, and ecological systems

¹ In this article, we use the analytic tradition of concept notation. If quoting the concept's usage, it will appear as "X" (e.g., Fisher's "diversity" is defined as...), when explicitly mentioned as a concept it will appear as X (e.g., the concept of diversity is...), and when implicitly mentioned as a concept it will appear as 'X' (e.g., 'heterogeneity' here describes...).

("ecosystems"). For example, empirical evidence shows that diversity improves academic performance (Gurin et al. 2004; Freeman and Huang 2015; Page 2014), because diverse individuals hold different values (Longino 1990; Harding 1991), situated knowledge (Haraway 1989), socio-gender locations (Code 2006), research styles and specialities (Gerson 2013) and conflicting theoretical scaffolds (Wimsatt and Griesemer 2007). There also are costs associated with diversity, including feelings of isolation and alienation leading to reduced academic achievements of minorities (Armor 1972; Holoién 2013) and unbridgeable disagreements among researchers that disintegrate research groups (Gerson 2013; Shavit and Silver, accepted for publication).

There also are societal costs of divergence between scientists and non-scientists.

Within the social realm, increased divergence from scientific worldviews may facilitate public manipulation by spreading ignorance – agnotology (Proctor and Schiebinger 2008) – and untrue and/or unjust environmental outcomes (Shrader-Frechette 2002). Within the scientific realm, divergence exempts scientists from responsibility for not assessing carefully enough social risks of generalizing their recommendations outside the laboratory, field, or model (Douglas 2009). Given the increasing science-society divergence, it is often non-experts who engage with the public – e.g., journalists teaching politicians about climate change or students teaching the underprivileged – which further widen the separation and may also silence local knowledge (Fricker 2007), e.g. by leading experienced mothers not to consider their comprehensive understanding and information as 'knowledge' compared to a young psychology student who never held a child, or depriving those living all their life near a spring to "know" their local flow rate compared to an

ecology student or governmental regulator who read published results taken at random from nearby streams (Shavit, Kolombus and Silver, accepted for publication).

Given the fine line between the costs and benefits of constructive and destructive dissimilarities, interrogating the most basic concepts and measurements of dissimilarity seems important and timely. This paper aims for a step in that direction.

2. Definitions of Dissimilarity

Fundamental to both diversity and heterogeneity is the concept of “variance” (Fisher 1918, 1925). Briefly, measurable properties (“variables”) of a group of individual entities (a “population” of cells, organisms etc.) are rarely identical. Rather, they will take on a range of values $\mathbf{y} = \{y_1, y_2, y_3, \dots y_n\}$, where the value of the variable measured for the i^{th} individual is denoted y_i . When graphed as a histogram (Tukey 1977), these values are distributed, with the most frequent values clustered around the most common one and rarer values towards the edges.

The average value of the distribution of the measured variables (its expected value $E(\mathbf{y})$ or its mean value \bar{y}), equals the sum of all the individual measurements divided by

the number of individuals, n : $\bar{y} = \sum_{i=1}^{i=n} \frac{y_i}{n}$. The variance, or “spread” of the distribution is

the sum of the squared differences between each individual measurement and the mean:

$\sigma^2 = \sum_{i=1}^{i=n} (y_i - \bar{y})^2$. The standard error of the mean $\left(\frac{\sqrt{\sigma^2}}{n}\right)$ provides intuitive estimates

of how variable the set of measurements is. Under reasonable assumptions, $\approx 63\%$ of the

measurements fall within ± 1 standard error of the mean, and $\approx 95\%$ fall within ± 2 standard errors of the mean.²

In statistics (and hence in nearly all the social and natural sciences), means and variances are characteristics of single populations (groups of measurements), but heterogeneity usually is a composite property of a group of measurements taken from more than one population. For example, the classic analysis of variance (ANOVA) developed by Fisher (1918) is used to determine if two or more populations differ in their average measured traits (e.g., height). A basic assumption of ANOVA is that the variances of the populations being compared are equal; this is referred to as “homogeneity of variance” or “homoskedasticity”. In contrast, if variances are unequal (heterogeneous or heteroskedastic), mathematical transformations of the data must be done to ensure that variances are homogeneous prior to comparing populations using ANOVA.³ Note that ‘heterogeneity’ here describes only the variance as a problem to overcome in order to allow a *common basis* for comparison. Throughout the rest of this article, however, the concept of heterogeneity describes entities within a collective. “Diversity”, if it is used at all in statistics, refers simply to describe a collection of datasets that describe a wide range of different, often incommensurate, variables.

In contrast, diversity is used widely in ecology (e.g., McGill et al. 2015) and the social sciences (e.g., Page 2011). Unlike variance or heterogeneity, diversity is not a simple, one-dimensional predicate. McGill et al. identified at least 15 different kinds of

² Ellison and Dennis (2010) provide a full discussion of the assumptions behind these estimates and calculation of associated confidence intervals.

³ See Gotelli and Ellison (2012) for details and another example of a “cost” of heterogeneity.

ecological diversity; differences among them reflect the number of variables or populations that are measured (one or more), the spatial scale of measurement (local or regional), and whether it is measured within or between populations. Unlike ‘variance’ or ‘heterogeneity’ – both of which are interpretable on their own – ‘diversity’ has little meaning to an ecologist unless it is associated with an object. For example, the concept of *alpha* diversity refers to the number of different species in a locality, the concept of *gamma* diversity to the number of different species in a region [a collection of localities], and *beta* diversity measures population change between localities.⁴

In the social sciences, Page (2011) makes similar distinctions between three kinds of diversity: (1) *variation*, or diversity within a type, referring to quantitative differences in a specific variable; (2) *diversity of types*, referring to qualitative differences between types; and (3) *diversity of composition*, or the way types are arranged. Page’s variation is directly analogous to an ecologist’s alpha diversity, and his diversity of types and diversity of composition are analogous to different dimensions of an ecologist’s beta diversity. Most social scientists use “diversity” as a catchall phrase not attached to any particular measured process (Page, personal communication), but we suggest that more attention should be paid to the dimensions of beta diversity.

Although ‘diversity’ appears to be used abstractly in common parlance and is implicitly assumed to mean something very similar to ‘heterogeneity’, when we examined deeply rooted linguistic intuitions of certain core examples, and tested these intuitions in large databases of linguistic usage, an interesting distinction between ‘diversity’ and

⁴ Each of these can be unweighted (i.e., simple counts of different species) or weighted by their abundance or sizes (Chao et al. 2014).

‘heterogeneity’ was revealed, with relevance for understanding and improving civil society and its institutions.

3. A Conceptual Tension Between Diversity and Heterogeneity

Whereas scientific language may seem indecisive or vague, artistic language can be precise and revealing. For example, Robert Frost’s *The Road Not Taken* beautifully highlights diverging dimensions of a difference (i.e., ‘diversity’), whereas the etymology of ‘heterogeneous’ implies something quite the opposite: an integration of multiple other (Gr.: *hetero*) kinds (Gr. *genus*) within a single whole.

We argue that attributing heterogeneity to something (e.g., a cell, computer, etc.) implies attributing an *integration* of mutual interactions among different entities that all belong to the same *collective*, whereas attributing diversity to a collection of objects or entities entails neither interactions nor a common collective.

An examination of English idiomatic constructions reveals clear distinctions in usage of diversity and heterogeneity. We would say that the parts of a cell or a clock are heterogeneous, but not that they are diverse. In contrast, we recognize a diverse collection of wall decorations or tools. There is an apparent semantic distinction here: cells and clocks are collectives whose functioning entails the integration of a number of interacting parts, whereas walls or garages function independently of the collection of items hanging on them. In other aspects of common usage, however, many objects in daily speech, including communities, populations, or universities, are called diverse or heterogeneous interchangeably.

The *Corpus of Contemporary American English* (henceforth: COCA; Davies 2008) provides a resource with which to examine common usage of diversity and heterogeneity in more detail. COCA contains more than 520 million words of texts, including scholarly

writing, fiction and nonfiction, newspapers and spoken recordings, and has tools to conduct complex searches for occurrences of words, phrases, parts of speech, other linguistic forms, and any combination thereof. Compilations of lists of co-occurrences (i.e., all types of words [adjectives, verbs, nouns, etc.] or specific words that appear near a target word) that can be used to infer intended meanings of predicates such as diverse or heterogeneous.

Sabar (2016) used COCA to infer motivations underlying regular co-occurrences of words. By identifying partial intersection of words that regularly co-occur more than expected by chance alone, Sabar identified *communicative strategies*: the choices of specific linguistic forms that best contribute to their intended message (e.g., “look” and “carefully” form the phrase “look carefully” that calls for visual attention). Thus, the generality of a communicative strategy that is evident in a particular example is established via a quantitative prediction of a non-random co-occurrence (“look” and “carefully” occur together and in sequence more frequently than expected by chance alone, and Sabar (2016) confirmed that “look” and “see” differ in meaning as a feature of attention by showing that “look” co-occurred more frequently with words such as “notice” than did “see”).

We searched COCA and the *Wikipedia Corpus* (Davies 2015) for frequencies of “diverse” and “heterogeneous” and tested our hypotheses regarding differences in meaning between them using chi-square tests for non-random frequencies. “Diverse” occurred 12-30 times more frequently than “heterogeneous” in the corpora. In line with our hypothesis, “homogeneous”, “collective”, “whole”, “integration” and “interaction” co-occurred significantly more frequently with “heterogeneous” than with “diverse” (improved prediction by, respectively, 58, 24, 8, 11, and 11%). Antonyms of these words (“single”,

“individuals”, “division”, “separation”) showed only random patterns of co-occurrence when they co-occurred at all (see tables 1-7 in the Appendix). A possible explanation for the latter findings is that while concepts of a collective whole seem to be more explicitly related to ‘heterogeneity’, words and meanings of singularity are relevant to both terms (in the case of heterogeneity they could relate both a single whole or to its parts). Nonetheless, it is evident that there is empirical support for our semantic intuition regarding ‘heterogeneity’ as interactions among diverse entities within a collective whole, and, perhaps more importantly, the empirical lack of a collectivist meaning for ‘diversity’.

The attribute of diversity does not correctly describe collective entities because its meaning and reference are much wider than the concept of heterogeneity. A heterogeneous entity may be composed physically of nothing more than diverse entities, but as a collective, it entails multiple direct and indirect interactions, and feedbacks, among these entities. All reproducing biological groups (genomes, cells, metapopulations, etc.) are heterogeneous in the collective sense. Hence, additional information that refers to internal interactive processes improves models of heterogeneous entities and systems (Wade 1978; Roughgarden, accepted for publication). Some human groups – e.g., families, football teams or kibbutzim – would best be described as heterogeneous, whereas others – e.g., people waiting to pay the cashier – would not (Shavit 2008). There may be grave costs associated with failing to identify the goals of certain human groups as diverse or heterogeneous, as the next section portrays.

4. Illustrating the Diversity-Heterogeneity Trade-Off

4.1 Moral costs

Many – perhaps most – readers of this essay would say that promoting diversity is a social good because it is a stepping-stone to heterogeneity and thus to social justice. Although we may not yet have achieved a just and heterogeneous society, we should nonetheless promote diversity as much as possible and not dwell on the semantic particularities of distinguishing the concepts of diversity from heterogeneity. We think this line of thinking is misleading, and that the continuous focus on racial, ethnic, or gender ‘alpha diversity’ (i.e., headcounts) and use of the results of these measurements as a sufficient basis for discourse and policy, creates a vicious circle that may hinder social change in many of our institutions, in particular in our schools, colleges, and universities.

For example, in *Brown v. Board of Education* (1954), the Supreme Court of the United States ruled that segregation of African-American and Caucasian students in schools violated the Equal Protection Clause of the U.S. Constitution. One outcome of this decision was transporting students of different racial backgrounds into different school districts (“busing”) to achieve diverse, “integrated” schools. This was intended to provide equal opportunities, academic aspirations, and achievements for all students and to improve relations among different races (Armor 1972). Unfortunately, according to some of its strongest supporters, busing did not improve academic aspirations or achievements (St. John 1975), sometimes decreased them and often worsened interracial relations: “integration ... enhances ideologies that promote racial segregation, and reduces opportunities for actual contact between the races.” (Armor 1972, 13).

In higher education, diversification is primarily done through “affirmative action”. Many scholars support affirmative action (e.g., Bowen and Bok 2000; Rothstein and Yoon

2008), but others have argued that it leads to similar or worse outcomes than would have occurred in its absence (e.g., Sander 2004; Sander and Taylor Jr. 2012). For example, between 1988 and 2007, faculty of color made up only 17% of total full-time faculty, and that there had been little change in this number since the 1980's (Turner, González, and Wood 2008). Similar findings have been reported for the number of earned PhDs (NSF 2013).

However one thinks about affirmative action, we suggest that in the interest of promoting social justice that institutions should not measure diversity alone – how many people of different backgrounds are found at a certain time and place – nor wait for it “to work its magic” and reduce injustice. Smith (2015) identifies three problems with current mechanisms for promoting diversity in higher education: (1) responding to calls to improve diversity reactively rather than proactively, often by producing an internal quantified response to an external standardized requirement; (2) failure to include people from the many interacting parts of a university – faculty, staff, students, etc. – in discussions about diversity; and (3) making diversification into a specific program rather than an integral institutional function and goal. All of these common methods of “working towards diversity” are problematic precisely because they increase diversity but reduce heterogeneity. They track and magnify difference and divergence rather than encourage and enhance mutual interaction among all different co-occurring identity groups.

A more positive approach was reported by Walton and Cohen (2011), who conducted a very brief intervention in one's sense of social belonging (SOB) to a selective, largely Caucasian, college. After three years, there was a significant increase in the GPA (grade point average) of African-American students relative to control groups. SOB is central to a

heterogeneous community as it is a psychological aspect of being a part of an integrated collective.

We suggest that a trade-off exists between tracking diversity and building heterogeneity, which may result in a vicious circle leading to blaming those afflicted with social inequality for their under-representation. Since we are better at measuring discrete variables such as grades and gender than at measuring interactions such as SOB and research cooperation, we invest more effort in creating changes we can easily track rather than those that demand more complex, “beta type”, measurements (e.g., institutional SOB, type of contacts with colleagues or task composition in the lab). As a result of neither measuring these latter dynamics nor investing in their visible change, alienation and lower academic achievements may persist among minority students and scholars (Syed, Azmitia, and Cooper 2011) even while their “diversity” increases. If this processes continues, a dangerous positive feedback may emerge, where not only will one’s self-image and achievements be worsened, but also his/her social identity comes out worse than before affirmative action took place.

4.2. Epistemic Benefits

Aiming for heterogeneity rather than diversity often has epistemic benefits. Human collectives – as well as individual agents – have a variety of epistemic perspectives (Shavit, Kolumbus and Silver, accepted for publication). These perspectives differ in multiple inter-related ways, involve different backgrounds and experiences, and vary in ways of perceiving, explaining, and evaluating information about the world. Perspectives direct our attention to track a wide range of phenomena, promote diverse models to explain them (Griesemer 2014) and encourage adaptive-reflection by employing “...a variety of social perspectives, often...by taking the perspective of others” (Bohman 2006, 180).

Information is distributed asymmetrically between agents, so that some of it is known in general, some exclusive to certain groups, and some idiosyncratic to specific individuals (Sunstein 2003; Anderson 2006; Solomon 2006a; Gerson 2013); lack of interaction keeps pieces of information latent.⁵ Diversity alone will not ensure that information is shared and provides fewer opportunities for agents to reflect on information that they can access only through interactions with others (Longino 2002; Tollefsen 2006).

Integrative working interaction across specialties – unlike the typical diverse-one-way adoption of ideas from one disciplinary to another – “includes coordinated efforts to pose and solve new research problems that can redefine specialty boundaries” (Gerson 2013, 516), and leads to developing new specialties. Tollefsen (2006) interweaves individual and collective knowledge in a way that demonstrates the benefits of epistemic heterogeneity. She suggested a framework of splitting a group that shares a common goal (e.g., works on a related set task or problems) into sub-groups; heterogeneity is manifested on an inter-sub-group level. Each sub-group is responsible for a different task, has its own sub-goals, and devises its own strategies and solutions. Mutual interactions result when the sub-groups return to the original group setting to present their suggestions and give feedback to other sub-groups. They encounter dissenting perspectives of out-groups and are forced to consider them and examine their own perspective closely. This self-scrutiny and actual encounters with critiques by other groups reveals problems, such as inaccuracies, leaps and gaps, and uncertainties, allowing the sub-groups and the integrated collective opportunities for self-correction (Tollefsen 2006).

⁵ There is an on-going discussion regarding the epistemic efficacy of deliberation, which is beyond the scope of this article.

Since all sub-groups are part of a larger community that shares a common goal, they both depend on other sub-groups and are depended upon by them. This framework is heterogeneous rather than diverse as the common goal and the inter-sub-group interactions serve to integrate the group. It also maintains differences, thus reducing the danger of group cohesiveness leading to unanimity and conformism, without promoting divergence. Such a framework increases the chances of achieving accurate results and obtaining a more just process of decision-making.

5. Conclusion

Diversity is not heterogeneity, and a continued focus on the former is not increasing the latter; instead, there is often a trade-off and tension between them. We illustrated how heterogeneity can better advance academic institutions and governance structures by integrating different people, identities, perspectives, and sources of information; it facilitates interactions among them, which have constructive epistemic and moral implications. Conversely, diversity alone often leads to divergence, is insufficient to resist social injustice and it misses epistemic opportunities that result from integrative working interactions. Institutions are often unaware of the diversity-heterogeneity tension or remain indifferent to it. They invest efforts in promoting diversity while neglecting heterogeneity, thus paying the costs of the trade-off and not reaping its benefits. Tracking alpha and disregarding beta diversity maintain this trade-off and obscures it. For moral and epistemic reasons we suggest noting this conceptual and practical difference and aiming for heterogeneity.

References

- Anderson, Elizabeth. 2006. "The Epistemology of Democracy." *Episteme* 3 (1-2): 8–22.
- Armor, David J. 1972. "The Evidence on Busing." *Public Interest* 28:90–126.
- Bohman, James. 2006. "Deliberative Democracy and the Epistemic Benefits of Diversity." *Episteme* 3 (3): 175–91.
- Bowen, William G., and Derek Bok. 2000. *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions*. Princeton, NJ: Princeton University Press.
- Chao, Anne, Nicholas J. Gotelli, T. C. Hsieh, Elizabeth L. Sander, K. H. Ma, Robert K. Colwell, and Aaron M. Ellison. 2014. "Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies." *Ecological Monographs* 84 (1): 45–67.
- Code, Lorraine. 2006. *Ecological Thinking: The Politics of Epistemic Location*. *Ecological Thinking: The Politics of Epistemic Location*. Oxford, UK: Oxford University Press.
- Davies, Mark. 2008. "The Corpus of Contemporary American English: 520 Million Words, 1990-Present." Accessed February 15. <http://corpus.byu.edu/coca/>.
- . 2015. "The Wikipedia Corpus: 4.6 Million Articles, 1.9 Billion Words." Adapted from Wikipedia. Accessed February 15. <http://corpus.byu.edu/wiki/>.
- Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.
- . 2015. "Politics and Science: Untangling Values, Ideologies, and Reasons." *The ANNALS of the American Academy of Political and Social Science* 658 (1): 296–306.
- Ellison, Aaron M., and Brian Dennis. 2010. "Paths to Statistical Fluency for Ecologist." *Frontiers in Ecology and the Environment* 8 (7): 362–70.
- Fisher, Robert A. 1918. "The Correlation between Relatives on the Supposition of Medelian Inheritance." *Transactions of the Royal Society of Edinburgh* 52:399–433.

- . 1925. *Statistical Methods for Research Workers. Biological Monographs and Manuals*. Edinburgh: Oliver and Boyd.
- Freeman, Richard B., and Wei Huang. 2015. “Collaborating with People like Me: Ethnic Co-Authorship within the US.” *Journal of Labor Economics* 33 (3(S1)): S289–318.
- Fricke, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford, NY: Oxford University Press.
- Frost, Robert. 1916. *Mountain Interval*. New York, NY: Henry Holt.
- Gerson, Elihu M. 2013. “Integration of Specialties: An Institutional and Organizational View.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:515–24.
- Gotelli, Nicholas J., and Aaron M. Ellison. 2012. *A Primer of Ecological Statistics. 2nd Edition*. Sunderland, MA: Sinauer Associates.
- Griesemer, James R. 2007. “Tracking Organic Processes: Representations and Research Styles in Classical Embryology and Genetics.” In *From Embryology to Evo-Devo*, ed. Manfred D. Laubichler and Jane Maienschein, 375–433. Cambridge, MA: MIT Press.
- . 2014. “Reproduction and the Scaffolded Development of Hybrids.” In *Developing Scaffolds in Evolution, and Cognition*, ed. Linnda R. Caporael, James R. Griesemer, and William C. Wimsatt, 23–55. Cambridge, MA: MIT Press.
- Griesemer, James R., and Michael J. Wade. 1988. “Laboratory Models, Causal Explanations and Group Selection.” *Biology and Philosophy* 3 (1): 67–96.
- Gurin, Patricia, Jeffrey S. Lehman, Earl Lewis, Eric L. with Dey, Sylvia Hurtado, and Gerald Gurin. 2004. *Defending Diversity: Affirmative Action at the University of Michigan*. Ann Arbor, MI: University of Michigan Press.

- Haraway, Donna. 1979. "The Biological Enterprise: Sex, Mind, and Profit from Human Engineering to Sociobiology." *Radical History Review* 20:206–37.
- . 1989. *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. New York, NY: Routledge.
- Harding, Sandra. 1991. *Whose Science? Whose Knowledge? Thinking from Women's Lives*. Ithaca, NY: Cornell University Press.
- Holoien, Deborah S. 2013. "Do Differences Make a Difference? The Effects of Diversity on Learning, Intergroup Outcomes, and Civic Engagement." University Report, The University of Princeton.
- Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.
- . 2002. *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.
- McGill, Brian J., Maria Dornelas, Nicholas J. Gotelli, and Anne E. Magurran. 2015. "Fifteen Forms of Biodiversity Trend in the Anthropocene." *Trends in Ecology and Evolution* 30 (2): 104–13.
- National Science Foundation, National Center for Science and Engineering Statistics. 2013. "Survey of Earned Doctorates, 1998–2013 [NSF Publication No. 15-304]." Accessed February 19. <http://www.nsf.gov/statistics/srvydoctorates/>.
- Page, Scott E. 2011. *Diversity and Complexity*. Princeton, NJ: Princeton University Press.
- . 2014. "Diversity without Silos: The Confluence of the Social and Scientific Teaching of Diversity." *Independent School Magazine* 73 (4): 27–30.
- Proctor, Robert N., and Londa Schiebinger, eds. 2008. *Agnotology: The Making and Unmaking of Ignorance*. Stanford, CA: Stanford University Press.
- Rothstein, Jesse, and Albert H. Yoon. 2008. "Affirmative Action in Law School Admissions: What Do Racial Preferences Do?" Working Paper 14276, National Bureau of Economic Research.

- Roughgarden, Joan. *Accepted for publication*. "Model of Holobiont Population Dynamics and Evolution: A Preliminary Sketch." In *Landscapes of Collectivity in the Life Sciences*, ed. Snait Gisis, Ehud Lamm, and Ayelet Shavit. Cambridge, MA: MIT Press.
- Sabar, Nadav. 2016. "A Meaning Hypothesis to Explain Speakers' Choice of the Sign Look." PhD diss., City University of New York.
- Sander, Richard H. 2004. "A Systematic Analysis of Affirmative Action in American Law Schools." *Stanford Law Review* 57 (367): 367–483.
- Sander, Richard, and Stuart Taylor Jr. 2012. *Mismatch: How Affirmative Action Hurts Students It's Intended to Help, and Why Universities Won't Admit It*. New York, NY: Basic Books.
- Shavit, Ayelet, Anat Kolumbus, and Yael Silver. *Accepted for publication*. "Epistemic Collectives, Heterogeneity and Injustice: The Case for Town Square Academia." In *Landscapes of Collectivity in the Life Sciences*, ed. Snait Gisis, Ehud Lamm, and Ayelet Shavit. Cambridge, MA: MIT Press.
- Shavit, Ayelet, and Yael Silver. *Accepted for publication*. "To Infinity and Beyond!" Inner Tensions in Global Knowledge- Infrastructures Promote Local and pro-Active 'location' Information." *Science and Technology Studies*.
- Shavit, Ayelet. 2008. *One for All? Facts and Values in the Debate over the Evolution of Altruism*. Jerusalem: Magness Press, in Hebrew.
- Shrader-Frechette, Kristin. 2002. *Environmental Justice: Creating Equality, Reclaiming Democracy*. Oxford, UK: Oxford University Press.
- Smith, Daryl G. 2015. *Diversity's Promise for Higher Education: Making It Work*. 2nd Edition. Baltimore, MD: Johns Hopkins University Press.

- Solomon, Miriam. 2006a. "Groupthink versus The Wisdom of Crowds: The Social Epistemology of Deliberation and Dissent." *The Southern Journal of Philosophy* 44 (1): 28–42.
- . 2006b. "Norms of Epistemic Diversity." *Episteme* 3 (1): 23–36.
- St. John, Nancy H. 1975. *School Desegregation: Outcomes for Children*. New York, NY: Wiley.
- Sunstein, Cass. 2003. *Why Societies Need Dissent*. Cambridge, MA: Harvard University Press.
- Syed, Moin, Margarita Azmitia, and Catherine R. Cooper. 2011. "Identity and Academic Success among Underrepresented Ethnic Minorities: An Interdisciplinary Review and Integration." *Journal of Social Issues* 67 (3): 442–68.
- Tollefsen, Deborah. 2006. "Group Deliberation, Social Cohesion, and Scientific Teamwork: Is There Room for Dissent?" *Episteme* 3 (1-2): 37–51.
- Tukey, John W. 1977. *Exploratory Data Analysis*. New York, NY: Addison-Wesley.
- Turner, Caroline Sotello Viernes, Juan Carlos González, and J. Luke Wood. 2008. "Faculty of Color in Academe: What 20 Years of Literature Tells Us." *Journal of Diversity in Higher Education* 1 (3): 139–68.
- Wade, Michael J. 1978. "A Critical Review of the Models of Group Selection." *The Quarterly Review of Biology* 53 (2): 101–14.
- Walton, Gregory M., and Geoffrey L. Cohen. 2011. "A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students." *Science* 331 (6023): 1447–51.
- Wimsatt, William C., and James R. Griesemer. 2007. "Reproduction Entrenchments to Scaffold Culture: The Central Role of Development in Cultural Evolution."

In *Integrating Evolution and Development: From Theory to Practice*, ed. Roger Sansom and Robert N. Brandon, 227–324. Cambridge, MA: MIT Press.

Appendix

Table 1. Wikipedia Corpus total target words occurrences.

Diverse	Heterogeneous
30967	1096

Table 2. Co-occurrences of “heterogeneous”/ ”diverse” with “interaction”. Hypothesis: “heterogeneous”-“interaction” > “diverse”-“interaction”.

	<i>Interaction present</i>		<i>Interaction absent</i>	
	N	%	N	%
Heterogeneous	11	18	1085	7
Diverse	49	82	30918	93
Total	60	100	32003	100

$P < .001$

Table 3. COCA total target words occurrences.

Diverse	Heterogeneous
16685	1305

Table 4. Co-occurrences of “heterogeneous”/ ”diverse” with “collective”. Hypothesis: “heterogeneous”- “collective” > “diverse”- “collective”.

	<i>Collective present</i>		<i>Collective absent</i>	
	N	%	N	%
Heterogeneous	5	31	1300	7
Diverse	11	69	16674	93
Total	16	100	17974	100

$P < .001$

Table 5. Co-occurrences of “heterogeneous”/ ”diverse” with “whole”. Hypothesis: “heterogeneous”- “whole” > “diverse”- “whole”.

	<i>Whole present</i>		<i>Whole absent</i>	
	N	%	N	%
Heterogeneous	7	15	1298	7
Diverse	40	85	16645	93
Total	47	100	17943	100

$P < .05$

Table 6. Co-occurrences of “heterogeneous”/ ”diverse” with “integration”. Hypothesis: “heterogeneous”- “integration” > “diverse”- “integration”.

	<i>Integration present</i>		<i>Integration absent</i>	
	N	%	N	%
Heterogeneous	6	18	1299	7
Diverse	28	82	16657	93
Total	34	100	17956	100

$P < .05$

Table 7. Co-occurrences of “heterogeneous”/ ”diverse” with “single”. Hypothesis: “heterogeneous”- “single” < “diverse”- “single”.

	<i>Single present</i>		<i>Single absent</i>	
	N	%	N	%
Diverse	77	97	16608	93
Heterogeneous	2	3	1303	7
Total	79	100	17911	100

$P > .05$

Levels of Reasons and Causal Explanation

Abstract

My starting points are the claims that explanations are answers to why-questions, and that to answer the question why some event E occurred one must provide reasons why E occurred. The idea that all explanations of events are causal then becomes the theory that the reasons why some event occurred are its causes. My main thesis in this paper is that many “counterexamples” to this theory turn on confusing two levels of reasons. We should distinguish the reasons why an event occurred (“first-level reasons”) from the reasons why those reasons *are* reasons (“second-level reasons”). An example that treats a second-level reason as a first-level reason will look like a counterexample if that second-level reason is not a cause. But second-level reasons need not be first-level reasons; nor (on my theory) need they be causes. Along the way I use the distinction between levels to diagnose the appeal of, and one main flaw in, the DN model of explanation.

1 A New Causal Theory of Explanation

It is obvious that some explanations of some phenomena speak of the causes of those phenomena. Simple examples come immediately to mind: the bridge collapsed because the wind reached a certain intensity, electrons flew off the metal because light shone on it. Much more controversial is the claim that *every* explanation of why some event happened must say something about the causes of that event. What's more, not only is it controversial whether this claim is true, it is also controversial how the claim should be understood. I have a new way of understanding the idea that all explanations of events invoke causes, one that, I think, is the most natural way to understand it. I also think that the idea, understood my way, is true (with one qualification¹), and can be defended against the repeated claim that there exist non-causal explanations.

My theory starts with the idea, which has been held by many others, that explanations are answers to why-questions.² A theory of explanation, then, should say what it takes for a proposition to be an answer to a why-question. Now one standard form answers to why-questions take is "P because Q": "The tide is high because the moon is overhead" answers "Why is the tide high?" But there is another form answers to why-questions can take. The other form is "A/The reason why P is that Q."³ Now because-answers and reasons-why answers are, in some sense, equivalent. "The tide is high because the moon is overhead" and "The reason why the tide is high is that moon is overhead" in some sense convey the same information. But I think that, for theoretical purposes, it is better to focus on reasons-answers. (I argue for this claim in (Skow 2016).)

A theory built around reasons-why answers will fill in the schema

¹See footnote 6.

²Among those who hold that explanations are answers to why-questions are Hempel (1965)—with some qualifications, Bromberger (1992), and Van Fraassen (1980).

³I ignore here the forms used to give "teleological" explanations; I extend my theory to cover teleological explanations in (Skow 2016).

1. A reason why P is that Q iff . . .

What should the claim that “explanations of events are causal” look like, if put into the form (1)? Let “P” hold the place for a sentence that describes the occurrence of an event. (I won’t try to say anything useful about which sentences do this.) Here is my proposal:⁴

(T) A reason why P is that Q if and only if the fact that Q is a cause of the fact that P.⁵

The same kinds of examples that lend credence to the idea that explanations of events are causal lend credence to its translation (T) into the language of reasons. The lighting of the fuse caused the bomb to go off; sure enough, it is also true that the reason why the bomb went off is that the fuse was lit. The electron’s passing through a magnetic field caused it to accelerate; sure enough, the reason why it accelerated is that it passed through a magnetic field.

On the other hand, the same examples philosophers have thought are counterexamples to the idea that explanations of events are causal also threaten to be counterexamples to (T).

A bunch of these examples, I think, are based on the same mistake. There is a distinction to be made between “levels” of reasons. The examples fail because they confuse the two levels.⁶ My aim in this paper is to introduce the distinction, and show how it can be used to defuse some examples. I will look, in particular, at Elliott Sober’s claim that equilibrium explanations are non-causal, and Marc Lange’s claim that “distinctively mathematical” explanations are non-causal (Sober 1981, Lange 2013).

⁴There are other theories of explanation that try to capture the idea that all explanations of events are causal—for example, (Salmon 1984) and (Lewis 1986). I do not have space here to explore the differences between their theories and mine.

⁵For stylistic convenience I sometimes speak of causation as a relation between facts, and sometimes as a relation between events. I remain neutral on which, if either, of these ways of speaking gets us closer to causation’s “fundamental nature.”

⁶I should say that there is one kind of counterexample that I think succeeds against (T): examples of “grounding” explanations. My true view is that every reason why a given event occurred is *either* a cause *or* a ground of its occurrence. But I will ignore grounding explanation in this paper.

2 Levels of Reasons

The distinction I want to introduce is that between

- a fact R being a reason why some event E occurred—then R is a “first-level” reason; and
- a fact F being a reason why R is a reason why E occurred—then F is a “second-level” reason, a reason why something else is a reason.

Reasons on the two different levels appear in answers to different why-questions. The first-level reasons are the facts that belong in the complete answer to the question *why E occurred*. The second-level reasons, on the other hand, belong in the answer to a different why-question: the question, concerning some reason R why E occurred, of *why R is a reason why E occurred*.

It is easy to come up with examples of first-level reasons. If I strike a match and, by striking it, cause it to light, then one reason why the match lit is that I struck it. What about an example of a second-level reason? We can find one by looking for the answer to the question of why the fact that I struck the match is a reason why the match lit. One answer (there are others) is: one reason why the fact that I struck the match is a reason why the match lit is that there was oxygen in the room at the time. In general, background conditions to a cause’s causing its effect are, I hold, reasons why the cause is a reason why its effect happened. (Background conditions are not, however, the only kind of second-level reason; more on this in a bit.)

3 Second-Level Reasons Need Not Be First-Level Reasons

Here is the thesis about levels of reasons that I will defend in this paper:

A fact can be a second-level reason without being a first-level reason. A fact F can be a reason why R is a reason why E happened, without F itself being a reason why E happened.

I say that F *need not* itself be a reason why E happened; I do not say that it *cannot*. The example I gave earlier shows that sometimes F *is* also a reason why E happened. The presence of oxygen,

besides being a reason why the striking of the match is a reason why the match lit, is also itself a reason why the match lit. But it is not always like this.

Here is an example in which a second-level reason is not also a first-level reason. Jill throws a rock at a window, Joan sticks out her mitt and catches the rock, and the window remains intact. The fact that Joan stuck out her mitt is a reason why the window remained intact. There is the first-level reason. *Why* is it a reason? The reason why it is a reason is that Jill threw a rock at the window. (You can test this with a counterfactual: if Jill hadn't thrown, certainly Joan's sticking out her mitt would not have been a reason why the window remained intact. The window wouldn't have "needed" Joan's help.) But this second-level reason is not also a first-level reason: that Jill threw a rock is *not* a reason why the window remained intact.⁷

In this case, the second-level reason that is not also a first-level reason is a fact that "corresponds" to the occurrence of an event: Jill's throwing of the rock. According to my theory (T), first-level reasons why events occur all correspond to events, since they are all causes. But not all second-level reasons are like the two examples we've seen so far (Jill's throw, the presence of oxygen); not all second-level reasons correspond to events.

In fact, I hold that laws of nature are second-level reasons that are not also first-level reasons. If I drop a rock from one meter above the ground, and it hits the ground at a speed of 4.4 m/s, the fact that I dropped it from one meter up is a reason why it hit the ground at 4.4 m/s. The law relating impact speed s to drop height d , namely $s = \sqrt{2dg}$ (assuming drag is negligible and d is small), is a second-level reason: it is a reason why my dropping the rock from one meter up is a reason why the rock was going 4.4 m/s when it landed. But it is not, in my view, also a first-level reason. It is not a reason why the rock is on the ground at 4.4 m/s.

Mentioning laws of nature probably brings to mind Carl Hempel's DN model of explanation, which says (I'm sure this is familiar) that an explanation of a fact F is a conjunction of facts that (i) entail F, and (ii) essentially contains a law among its conjuncts (Hempel 1965).

⁷This is also the kind of example many take to show that causation is not transitive; see for example (Hitchcock 2001).

Hempel's theory is not framed as a theory of the reasons why facts obtain, but it is natural to interpret it as committed to the thesis that whenever there are any reasons why some fact obtains, at least one of the reasons is a law of nature. I, along with many others, reject Hempel's theory, but I have a new diagnosis of where it goes wrong. Its mistake is to take certain second-level reasons, laws of nature, to also be first-level reasons.

I asserted without argument that laws are second-level reasons; but this is a natural view to have, on certain approaches to causation. One approach to causation takes laws to be central: whenever you have a cause and effect C and E, there are some laws connecting C to E—and C is a cause of E *because of* those connecting laws.⁸ But that is just to say that whenever C is a cause of E, some law is a reason why C is a cause of E. Now I hold that when some fact F is a reason why C is a cause of E, then F is also a reason why C is a reason why E happened. So it follows from this theory of causation that laws are second-level reasons. If you start here, and in addition think that second-level reasons are always also first-level reasons, you head toward the characteristic thesis of the DN model, the thesis that among the reasons why some event happens is always at least one law. But this line of thought is fallacious, because second-level reasons need not be first-level reasons; and, on my view, laws that are second-level reasons are never first-level reasons.

I admit that I have given no direct argument that laws are not first-level reasons. I'd like to put the burden on the other side: why think they are? They are certainly second-level reasons: they are certainly reasons why causes are reasons why their effects happen. But as the Joan and Jill example shows, second-level reasons are not always first-level reasons. So why think they are in the case of laws? Certainly we have a sense that laws are "explaining something"; my view captures this sense, by assigning them the role of explaining why causes explain their effects. Why isn't that enough?

⁸Hempel endorses something like this idea about causation; see (Hempel 1965: 349). It has, of course, had many other defenders.

4 How The Levels Can Get Confused

I said that the flaw in the DN model is that it mis-classifies laws, which are second-level reasons, as first-level reasons. I also sketched an argument (with a false premise) that leads to this mis-classification: “laws are second-level reasons, and second-level reasons are always first-level reasons, so laws are also first-level reasons.” But I’m not saying that Hempel or anyone else ever entertained this argument explicitly. Is there anything else to be said about how and why supporters of the DN model might have come to mis-classify laws as first-level reasons?

Yes, there is. Pragmatic effects, effects of the rules of conversation on information exchange, can produce “data” that misleadingly suggest that laws are first-level reasons.

The reasons why an event happened are the parts of the answer to the question of why it happened. So if we come across a conversation in which one person asks “Why did E happen?,” and another person answers this question by citing some fact F; and if that answer strikes us as correct; then we have some good evidence that F really is a reason why E happened.

Some of the evidence that laws are (first-level) reasons why events happen appears to fit this pattern (but I will argue it does not). Imagine someone walks into the room just as the rock hits the ground at 4.4 m/s, and she sees that it hit at this speed (maybe the rock fell onto a device that measures impact speeds). A curious person, she asks me why it hit the ground at 4.4 m/s. I respond,

Well, I dropped it from one meter up, and impact speed s is related to drop height d by the law $s = \sqrt{2dg}$ (and of course $\sqrt{2 \cdot 1 \cdot 9.8} \approx 4.4$).

Haven’t I answered her question? And doesn’t the law that $s = \sqrt{2dg}$ appear in my answer? If so, then the law is a reason why the rock hit the ground at 4.4 m/s—isn’t it?

If the answers to these questions are “yes, yes, and yes,” then, at least in some cases, a law is a reason why an event occurred. It’s not hard to get from this conclusion to the claim (characteristic of the DN model) that this is so in *all* cases, and that when someone answers a

why-question *without* mentioning a law, her answer is incomplete.⁹

But the answers to these questions are not “yes, yes, and yes.” To explain what I think is going on I need to introduce another distinction: the distinction between a *good response* to a question and an *answer* to a question. If someone asks a question, obviously one good way to respond is to answer the question. But not every good response is an answer.

A simple example suffices to establish this. Sally asks whether Caleb is coming to the party. I know he’s supposed to go to the party. I respond by saying “He’s sick.” This is a good response. But it is not an answer. The only two possible answers are “yes (he’s coming)” and “no (he’s not coming).” I didn’t say either of those things.

There is a theoretical reason why we should expect there to be good responses that are not answers. The notion of an answer is a semantic one. The relation between a proposition and a question, in virtue of which that proposition is an answer to that question, is a semantic relation. But the notion of a good response is a pragmatic one. Whether a response to a question is good is a matter of what a cooperative speaker should say. In some circumstances, a cooperative speaker should respond to a question by doing something other than, or something more than, answering the question. In the simple example, I know that if I just answer the question by saying “no,” then Sally will immediately ask me why he’s not coming. Since I can foresee that she’ll ask that, and since I know the answer to this question too, I respond to her explicit question not by answering it, but by answering the expected follow-up question. It is okay in this case not to explicitly answer the question she asked, because what I do say, my answer to the expected follow-up, conversationally implies that the answer to her explicit question is no.

I did not, however, need to be so indirect. I could have responded by answering both questions. I could have said, “no, he’s sick.” Here my response is good, but again it contains information that is not part of the answer to the question she explicitly asked. What keeps it from being a bad response is that the additional information is relevant to the topic of our

⁹This “incompleteness” defense is most fully developed by Railton (1981). For one thorough argument against it, see (Woodward 2003: chapter 4).

conversation; and it is relevant because, though it is not an answer to her question, it is an answer to an expected follow-up question.

I think the same thing is going on in the dropped rock example. I responded to the question by saying

Well, I dropped it from one meter up, and impact speed s is related to drop height d by the law $s = \sqrt{2dg}$.

My response is a good one, but (as we've seen) it does not follow that every part of my response is part of an answer to the question asked. In my view, the first part of my response—"I dropped it from one meter"—is an answer to the explicit question ("why did the rock hit the ground at 4.4 m/s?"), but the second part, the law, is not; it, instead, is an answer to an unasked follow-up why-question, a follow-up question I can anticipate would be asked immediately if I only answered the explicit question. The follow-up is, of course, why is the fact that I dropped it from one meter up a reason why it hit the ground at 4.4 m/s?

In summary: it is often a good thing to include a second-level reason in a response to the question why some event happened; but the fact that this is good thing to do is compatible with that second-level reason not being a reason why that event happened.

5 Equilibrium Explanations

I now have two distinctions: that between first- and second-level reasons, and that between a good response to a why-question an answer to a why-questions. The two together provide the key to defusing many problem cases for (T), the thesis that the reasons why something happened are its causes.

Elliott Sober argued that equilibrium explanations are not causal explanations. His main example of an equilibrium explanation was R. A. Fisher's answer to the question of why the ratio of males to females in the current adult human population is very close to 1:1 (Fisher 1931). "The main idea" of Fisher's answer, Sober reports, "is that if a population ever departs from

equal numbers of males and females, there will be a reproductive advantage favoring parental pairs that overproduce the minority sex. A 1:1 ratio will be the resulting equilibrium point” (201). Parents who overproduce the minority sex are likely to have more grandchildren. So if males outnumber females in the population, the fitter trait is to be disposed to have more female children than male; being the fitter trait, this disposition should increase in frequency, with the result that the sex ratio is pushed from male-biased toward equality. The opposite happens if females outnumber males. Now Sober claims that this is not a causal explanation, since

a causal explanation...would presumably describe some earlier state of the population and the evolutionary forces that moved the population to its present configuration...Where causal explanation shows how the event to be explained was in fact produced, equilibrium explanation shows how the event would have occurred regardless of which of a variety of causal scenarios actually transpired. (202)

In other words: Fisher’s explanation does not say, for example, that the sex ratio in the year 1000 was such-and-such, and that this caused the sex ratio in the year 1100 to be such-and-such, and so on. Instead it consists of a bunch of conditional facts: for each year in the sufficiently distant past, if the sex-ratio in that year had had any “non-extreme” value (non-extreme meaning not all males or females), then the sex ratio today still would have been 1:1.

The first thing I want to say is that Sober makes a claim about what the causes of the current sex ratio are that I reject. He thinks that the only relevant causes of the fact that the sex ratio is currently 1:1 are facts of the form *the sex ratio at time T is m:n*. I’m with those who reject this claim. The fact that the sex ratio in 1000 was m:n is “too specific” to be a cause of the current sex ratio. There is a less specific fact, the fact that the percentage of males in 1000 was not 0 or 100%, that is as well placed to be the cause. The less specific fact is “better proportioned” to the effect than the more specific one; so it gets to be the cause.¹⁰

¹⁰A “proportionality requirement” on causation is defended in Yablo (1992) and Strevens (2008). The claim that examples of explanations that, like Fisher’s, abstract away from the nitty-

My disagreement with Sober might not seem to help much. Isn't Fisher's explanation still a counterexample to (T)? Even if the cause of the current sex ratio is that the sex ratio in the past was never extreme, Fisher's explanation doesn't cite this cause either; his explanation instead contains a bunch of other facts, namely the conditional facts described earlier. Doesn't it follow that these conditional facts, which are not causes, are reasons why the sex ratio is 1:1, and thus that (T) is false?

I deny that those conditional facts that Fisher offers up are reasons why the sex ratio is 1:1. But I can't just say this; for when Fisher offered those facts up in response to the question of why the sex ratio of 1:1, everyone celebrated his response, they did not reject it. How can his response be something to celebrate, if it didn't answer the question?

The distinctions I introduced earlier show why. Fisher's response was something to celebrate, because it was a *good response to the question*. But it can be a good response without containing an answer; in fact that's exactly what I think is going on.

I think that the reason why the sex ratio is now 1:1 is that the sex ratio in the past was never extreme. But this is not something anyone would believe, or even be able to come to know, without an accompanying answer to the question of *why* that is the reason. So a good response to the question of why the sex ratio is now 1:1 must include an answer to the question of why the fact that the sex ratio was never extreme in the past is a reason why it is 1:1 now. And *that's* the question that the conditionals in Fisher's response constitute an answer to. Those conditional facts are second-level reasons why some other fact is a reason why the sex ratio is 1:1.

gritty details of the causal process that produced the event being explained count as non-causal is repeated by Batterman in, for example, (Batterman 2000: 28) and (2010: 2). Batterman assumes that abstracting away from the details takes you away from the causes; but the proportionality requirement shows that in some cases at least this is not so. Less specific facts may be better proportioned to an effect than more specific ones.

6 “Distinctively Mathematical” Explanations

Marc Lange has recently described a class of explanations that he calls distinctively mathematical explanations, and argued that they are not causal explanations (Lange 2013). My interest is not in whether his examples qualify as non-causal by his criteria, but in whether they are counterexamples to (T). Here is one of the examples:¹¹

Why did a given person [say, Jones] on a given occasion not succeed in crossing all of the bridges of Königsberg exactly once (while remaining always on land or on a bridge rather than in a boat, for instance, and while crossing any bridge completely once having begun to cross it)?...[Because] in the bridge arrangement, considered as a network, it is not the case that either every vertex or every vertex but two is touched by an even number of edges. Any successful bridge-crosser would have to enter a given vertex exactly as many times as she leaves it unless that vertex is the start or the end of her trip. So among the vertices, either none (if the trip starts and ends at the same vertex) or two could touch an odd number of edges (488-89).

Here is what Lange says about why explanations like this one not causal explanations:

these explanations explain not by describing the world’s causal structure, but roughly by revealing that the explanandum is more necessary than ordinary causal laws are (491).

There is definitely something right, and deep, in what Lange says. But I do not think that his examples are counterexamples to (T).

Let P be the property of bridge-arrangements that a bridge-arrangement has if and only if either every land-mass or every land-mass but two is met by an even number of bridges. The (supposed) answer to the question of why Jones failed that Lange presents boils down to this:

¹¹This example is also discussed in detail by (Pincock 2007).

(2) The bridges of Königsberg lacked P; and, necessarily, if a bridge arrangement lacks P, then no one can cross all the bridges exactly once.¹²

Now if (2) really is the answer to the question, then my theory is false. So is (2) the answer? There are two parts to (2). First is the fact that the bridges lacked P. Now it is no problem for my theory to recognize that this fact is a reason why Jones failed. For this fact is certainly a cause of his failure. The challenge to my theory comes if the second fact in (2) is a reason why Jones failed. For the second fact, that necessarily, no one can cross all the bridges exactly once, if the bridges lack P, cannot be a cause of Jones' failure.

I want to say the same thing about this example that I've said about the others. (2), I maintain, is not an answer to the question of why Jones failed. (2) contains an answer *as a part*—the fact that the bridges lacked P. But it has another part, the necessary truth, that is not part of the answer. How is this compatible with the evident fact that (2) is a really good thing to say in response to the question of why Jones failed? Because the part of (2) that is not an answer to this question *is* an answer to an obvious follow-up why-question, namely, why is it that the bridges' lacking P is the reason why Jones failed?

Lange's diagnosis of this example, and the others he discusses, is quite sophisticated, and I don't have the space here to go in to all the things he says about them. Let me at least, however, mention one further thing he says. At one point he writes, "Even if [these examples] happen to appeal to causes, they do not appeal to them as causes...any connection they may invoke between a cause and the explanandum holds not by virtue of an ordinary contingent law of nature, but typically by mathematical necessity" (496). I am quite taken by this idea that an answer to a why-question might appeal to causes but not appeal to them *as* causes. What might this mean, in terms of reasons why? Here is a natural suggestion: maybe in some cases a cause is a reason why its effect happened, but it is false that the *reason why* the cause is a reason why its effect happened is that it is a cause. The suggestion continues: cases like that are examples

¹²I'm going to take Lange's qualifications about always remaining on land etc. as given.

of “non-causal explanations.”

I think the suggestion is plausible: if there truly are cases like that, they should be counterexamples to my theory. They are not, however, counterexamples to my theory as stated. I should amend my theory to make it more vulnerable:

(T2) A reason why P is that Q if and only if (i) the fact that Q is a cause of the fact that P, and (ii) the reason why the fact that Q is a reason why P is that the fact that Q is a cause of the fact that P.

Now the question is whether the Königsberg example, or any other example, is a counterexample to (T2). I have a lot of thoughts about this, but can only be brief here. Lange’s idea is that since the “connection” between the bridges’ lacking P, and Jones’ failure, is secured by a mathematical truth (a theorem of graph theory), the bridges’ lacking P, while a reason, is not a reason because it is a cause. I reject this claim. Even if the connection is secured by a mathematical truth, the cause is still a reason because it is a cause. This assertion requires defense, but I don’t have the space to defend it here.

7 Conclusion

In this paper I have presented a new causal theory of explanation that says that the reasons why an event occurred are its causes. I also drew two distinctions: that between the reasons why E happened, and the reasons why those reasons are reasons; and that between an answer to a why-question, and a good response to a why-question. I used these distinctions to defend the theory against the claim that equilibrium explanations and distinctively mathematical explanations are non-causal; and I believe the distinctions can be used to defend it against a wide variety of other examples.

References

- Batterman, Robert (2000). "Multiple Realizability and Universality." *British Journal for the Philosophy of Science* 51: 115-45.
- (2010). "On the Explanatory Role of Mathematics in Empirical Science." *British Journal for Philosophy of Science* 61: 1-25.
- Bromberger, Sylvain (1992). *On What We Know We Don't Know*. The University of Chicago Press and CSLI.
- Fisher, R. (1931). *The Genetical Theory of Natural Selection*. Dover.
- Hempel, Carl (1965). "Aspects of Scientific Explanation." In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, 331-496.
- Hitchcock, Christopher (2001). "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy* 98: 273-299.
- Lange, Marc (2013). "What Makes a Scientific Explanation Distinctively Mathematical?" *British Journal for the Philosophy of Science* 64: 485-511.
- Lewis, David (1986). "Causal Explanation." In *Philosophical Papers, Volume II*. Oxford University Press.
- Pincock, Christopher (2007). "A Role for Mathematics in the Physical Sciences." *Nous* 41: 253-75.
- Railton, Peter (1981). "Probability, Explanation, and Information." *Synthese* 48: 233-256.
- Salmon, Wesley (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.
- Skow, Bradford (2016). *Reasons Why*. Oxford University Press.
- Sober, Elliott (1983). "Equilibrium Explanation." *Philosophical Studies* 43: 201-10.
- Strevens, Michael (2008). *Depth*. Harvard University Press.
- Van Fraassen, Bas C. (1980). *The Scientific Image*. Oxford University Press.
- Woodward, James (2003). *Making Things Happen*. Oxford University Press.

Yablo, Stephen (1992). "Mental Causation." *The Philosophical Review* 101: 245-80.

In Defense of the Actual Metaphysics of Race

Abstract. In a recent paper, David Ludwig (2015, 244) argues that “the new metaphysics of race” is “based on a confusion of metaphysical and normative classificatory issues.” Ludwig defends his thesis by arguing that the new metaphysics of race is non-substantive according to three notions of non-substantive metaphysics from contemporary metametaphysics. However, I show that Ludwig’s argument is an irrelevant critique of actual metaphysics of race. One interesting result is that actual metaphysics of race is more akin to the metaphysics done in philosophy of science than mainstream analytic metaphysics.

1. Introduction

In David Ludwig’s (2015, 44) recent article “Against the New Metaphysics of Race,” he argues for the provocative thesis that “the new metaphysics of race” is “based on a confusion of metaphysical and normative classificatory issues.” Furthermore, to continue to engage in such a “methodologically dubious metaphysics of race” is, in Ludwig’s (2015, 262) opinion, “a bad idea.” Key to Ludwig’s critique is that he defines “metaphysicians of race” as “committed to the ideal of one fundamental ontology of race,” much like other metaphysicians engaged in mainstream analytic metaphysics (Ludwig 2015, 245). Furthermore, for Ludwig, “the new metaphysics of race” consists of disputes about “one fundamental ontology of race” (Ludwig 2015, 245). In his critique, Ludwig focuses on two debates in the new metaphysics of race.

The first is the debate about whether races exist according to the one fundamental meaning of ‘race’ in current, ordinary English in the United States (Ludwig 2015, 257). I’ll call this *the US race debate**.¹ According to Ludwig (2015, 251, 253, 256, 260), some interlocutors

¹ The asterisk is intentional. I’m calling this debate ‘the US race debate*’ because I think Ludwig has changed the focus of the relevant debate. I borrow the convention of using an asterisk to flag when the meaning of a term has been changed from Joshua Glasgow (2009, 140).

in the US race debate* are Anthony Appiah, Joshua Glasgow, Michael Hardimon, Sally Haslanger, Quayshawn Spencer, and Naomi Zack.

The second debate in the new metaphysics of race is about whether humans have races according to the one fundamental meaning of ‘race’ in the life sciences (Ludwig 2015, 254). I will call this *the biological race debate**. Ludwig (2015, 251, 253, 259) claims that, among others, the interlocutors of the biological race debate* are Robin Andreasen, Bernard Boxill, A.W.F. Edwards, Adam Hochman, Jonathan Kaplan, Koffi Maglo, Armand Leroi, Massimo Pigliucci, Neven Sesardic, and Alan Templeton.

Ludwig defends his thesis using an argument premised on the claim that the new metaphysics of race is non-substantive according to three notions of non-substantive metaphysics from contemporary metametaphysics: one from Eli Hirsch, one inspired from Theodore Sider, and one from Ludwig himself. The relevant background here is that recent metametaphysics has been preoccupied with what constitutes a “substantive” metaphysical dispute, which, roughly, is a dispute that is *really* about metaphysics as opposed to some other topic, like how we use language (Hirsch 2005, 67).

While I agree with Ludwig that to engage in a metaphysics of race that confuses metaphysical and normative classificatory issues is a bad idea, and while I think that the new metaphysics of race (as Ludwig defines it) might be based on such a confusion, I will show that the work that *actual* metaphysicians of race are doing involves no such confusion. In other words, the point of this paper is show that Ludwig’s argument is an irrelevant critique of the actual metaphysics of race.

For clarity, by ‘actual metaphysicians of race’, I’m talking about the same group of scholars that Ludwig is talking about in his critique, and by ‘actual metaphysics of race’ I’m

talking about the same body of work that Ludwig is talking about in his critique.² However, unlike Ludwig (2015, 245), I will not require actual metaphysicians of race or actual metaphysics of race to be “committed to the ideal of one fundamental ontology of race,” even with respect to a particular linguistic context.

I will begin by clarifying Ludwig’s argument and his defense of each premise. Second, I will show that even if Ludwig’s argument is a good critique of the new metaphysics of race, it’s irrelevant to the actual metaphysics of race. Finally, I will provide closing remarks where, among other things, I will clarify how the actual metaphysics of race is more akin to the metaphysics done in the philosophy of science than mainstream analytic metaphysics. As for objections, I will respond to them along the way.

2. Ludwig’s Argument and Its Defense

2.1 The Basic Argument

Though Ludwig does not state his argument explicitly, a charitable reconstruction of it is below:

- (1) If the new metaphysics of race is non-substantive, then it is based on a confusion of metaphysical and normative classificatory issues.
- (2) The new metaphysics of race is non-substantive.
- (3) So, the new metaphysics of race is based on a confusion of metaphysical and normative classificatory issues.

² For instance, like Ludwig (2015, 244), I consider Joshua Glasgow to be an actual metaphysician of race, and, like Ludwig (2015, 263), I consider Glasgow’s actual metaphysics of race to consist of work like his book *A Theory of Race* and his article “On the New Biology of Race.”

Ludwig states (3) as his thesis in the first paragraph of his opening remarks.³ Ludwig states (2) in his opening remarks as well and at several points throughout his paper.⁴ Ludwig also treats (2) as a reason for adopting (3).⁵ However, since there is a logical gap between (2) and (3), it's charitable to add (1) as a suppressed premise.⁶

2.2 Ludwig's Defense of His Premises

Though Ludwig takes the truth of (1) for granted, he offers three, in-depth defenses of (2) that utilize three different notions of non-substantive metaphysics. Ludwig's first defense of (2) is the following:

- (4) The new metaphysics of race is substantive only if there is exactly one allowable and fundamental ontology of race for each of its race debates.
- (5) If there is a plurality of legitimate biological subdivisions below the species level or a plurality of equally allowable specifications of 'race' for each race debate in the new metaphysics of race, then there is a plurality of allowable ontologies of race for each race debate in the new metaphysics of race.
- (6) The antecedent of (5) is true.
- (7) So, it's not the case that the new metaphysics of race is substantive.

Ludwig claims (4) in section 3.1 and justifies his constraint on substantive metaphysics from how he defines 'a metaphysics of *x*.' For Ludwig (2015, 245, 251), a project on the

³ See Ludwig (2015, 244).

⁴ See Ludwig (2015, 245, 260-262).

⁵ See, especially, sections 3.1-3.3 and 4 in Ludwig (2015).

⁶ [removed for blind review]

“metaphysics of x ” assumes that metaphysicians of x are committed to “one fundamental ontology” of x that rules out “a plurality of equally allowable ontologies” of x , at least for the relevant linguistic context.⁷ Since a substantive metaphysics of x must at least be a metaphysics of x , it follows that a substantive metaphysics of x requires exactly one allowable and fundamental ontology of x . Substituting ‘race’ for ‘ x ’ gives us (4).

As for (5), Ludwig states that the first disjunct of (5)’s antecedent leads to (5)’s consequent in section 2. Here Ludwig (2015, 247) follows Kaplan and Winther (2013) in arguing that if there is a plurality of equally legitimate but distinct ways of subdividing species into “legitimate biological kinds,” then “[e]mpirical evidence underdetermines the ontological status of race,” which in turn, permits a plurality of allowable ontologies of race (Ludwig 2015, 246-247). In particular, Ludwig (2015, 245, 247-249) argues that “both racial realism and antirealism” are allowable ontologies of race given different equally legitimate ways of subdividing a species, and even in the same race debate. An example is how Zack (2002) uses the fact that humans have no subspecies to defend racial anti-realism in the US race debate*, while Spencer (2014) uses the fact that humans have a population subdivision that matches the current US census racial scheme to defend racial realism in the same race debate.

Ludwig states that the second disjunct of (5)’s antecedent leads to (5)’s consequent in section 3.1. In his words, “If there is a plurality of equally allowable specifications of ‘race’, there is also a plurality of equally allowable ontologies of race” (Ludwig 2015, 251). Interestingly, Ludwig never defends this assertion because he takes it to be obviously true.

⁷ See Ludwig (2015, 251) for (4) and see Ludwig (2015, 245) for Ludwig’s view on the metaphysics of x .

Next, Ludwig defends (6) by defending the truth of each disjunct in the antecedent of (5). As for the first disjunct, Ludwig (2015, 246-247) argues that there is a plurality of legitimate biological divisions below the species level (e.g. population subdivisions, monophyletic levels, subspecies, etc.) because, first, legitimate biological kinds are *interest dependent*, and, second, there is a plurality of “explanatory interests” among biologists in different research contexts (e.g. population genetics, phylogenetic systematics, etc.). As for the second disjunct, Ludwig reaches it by making an induction from what’s going on in the two most popular race debates in the new metaphysics of race: which are the US race debate* and the biological race debate*.

Ludwig (2015, 254) argues that there is a plurality of equally allowable specifications of ‘race’ in the biological race debate* since biologists in different research programs use ‘race’ in different ways that suit their needs. For instance, Ludwig (2015, 254) points out that ‘race’ is often used as a synonym for ‘subspecies’ in systematic biology, but often used as a synonym for ‘ecotype’ in ecology. As for the US race debate*, Ludwig takes a more circuitous route to the conclusion that there is a plurality of equally allowable specifications of ‘race’ in that debate. First, Ludwig (2015, 255) appeals to Glasgow et al.’s (2009) empirical research on how Americans use ‘race’ to argue that ‘race’ is “polysemous” in the current US. Next, Ludwig (2015, 257-258) argues that the *context* for the US race debate* has not been “sufficiently specified” to narrow the debate to “exactly one fundamental candidate meaning of ‘race’ in the United States.” Hence, according to Ludwig, from induction, the second disjunct of (6) holds as well.

Ludwig’s second defense of (2) utilizes Hirsch’s notion of non-substantive metaphysics. The second defense is below:

- (8) A dispute is merely verbal if each side can plausibly interpret the other

side as speaking a language in which the latter's asserted sentences are true.

- (9) A dispute is non-substantive if it is merely verbal.
- (10) Each side can plausibly interpret the other side as speaking a language in which the latter's asserted sentences are true in the new metaphysics of race.
- (11) Thus, the new metaphysics of race is non-substantive.

(8) is a direct quote from Ludwig (2015, 259), which is itself a summary of Hirsch's (2005; 2008) view on non-substantive metaphysics.

Hirsch defends his distinction between merely verbal disputes and ones that aren't with several examples from the history of science and philosophy. For instance, Hirsch (2005, 73) shows that the dispute among classical physicists about whether a projectile's final velocity is equal to its initial velocity on Earth was not a merely verbal dispute because physicists on both sides could not charitably interpret the other side's assertions as true. In other words, both sides were using the same meanings of 'projectile', 'velocity', 'Earth', etc., and what they disagreed about were the laws of motion. In contrast, Hirsch (2008, 407-408) shows that the dispute between John Locke and Joseph Butler about whether a tree can survive a change in its parts was merely verbal since either side could charitably interpret the other side's assertions as true using the other's meaning of 'identity'. In short, a merely verbal dispute for Hirsch is one where the disputants are either talking past one another or merely arguing about how we do (or should) use language.

As for (9), we can infer that it's a premise from how Ludwig (2015, 259-260) uses 'merely verbal' and 'nonsubstantive' at this point in his paper. Furthermore, Ludwig's

vocabulary here is uncontroversial since it's the same vocabulary that Hirsch (2005, 67) uses.

As for (10), Ludwig endorses it when he says the following:

Realists like Andreasen, Edwards, Leroi, Sesardic, and Spencer can interpret antirealists as speaking the truth in a language in which 'race' refers to subspecies, populations with visible traits that mark relevant biological differences, populations with cognitive differences, and so on. Antirealists like Glasgow, Lewontin, Hochman, Maglo, and Zack can interpret realists as speaking the truth in a language in which 'race' refers to genetic clusters, patterns of mating, clades, and so on (Ludwig 2015, 259-260).

Finally, Ludwig defends (2) in a third way using his interpretation of Sider's notion of non-substantive metaphysics. Ludwig's third defense of (2) is below:

- (12) A dispute about an expression *E* is non-substantive if its disputants are endorsing multiple, equally joint-carving candidate meanings for *E*.
- (13) The new metaphysics of race is a dispute that is non-substantive according to (12).
- (14) The new metaphysics of race is non-substantive.

(12) is directly from Ludwig (2015, 261), and is a rough summary of Sider's (2011, 46-49) view of non-substantive metaphysics. Sider defends the non-joint-carving condition in his definition of 'non-substantivity' from his stipulation of what metaphysics is about.

For Sider (2011, vii) the "central task" of metaphysics is "to discern the ultimate or fundamental reality underlying the appearances." We are supposed to describe this reality using a privileged language, so-called Ontologese, which is privileged exactly because all of its expressions (e.g. terms, quantifiers, etc.) are "joint-carving," which means that they carve out the

world's fundamental structure (Sider 2011, vii).⁸ So, naturally, when we find that one or more of the expressions that we've used to formulate a question Q does not have exactly one, best joint-carving meaning, it's likely that a debate about Q is not about the fundamental structure of the world, and thus, is not a substantive metaphysical debate in Sider's sense.

With that said, it's important to note that Ludwig's summary of Sider is rough, and does not reflect Sider's (2011, 49) "revised" definition of a non-substantive dispute. What Ludwig presents is Sider's unrefined view, which occurs at the beginning of section 4.2 in chapter 4 of Sider's *Writing the Book of the World*. However, later on in section 4.2, after Sider considers multiple problems with his unrefined view, he settles on what he calls his "revised" definition.⁹ Nevertheless, since Ludwig uses Sider's unrefined notion of non-substantivity in his critique, that's what I'll focus on as well. However, for clarity, I'll say that (12) expresses *Sider-style non-substantivity* as opposed to Siderian non-substantivity.

In any case, Ludwig (2015, 261) asserts and defends (13) when he says that Spencer's, Leroi's, Pigliucci's, and Hochman's biological definitions of 'race' are all "equally joint-carving candidates" for 'race' because they are all "objective ways of distinguishing between populations below the species level." Furthermore, Ludwig (2015, 261-262) bolsters his support for (13) when he says that Hardimon's, Glasgow's, Feldman and Lewontin's, and Appiah's biological definitions of 'race' are also equally joint-carving candidates for 'race' because they are all "non-joint-carving" meanings.

3. Why Ludwig's Argument is an Irrelevant Critique of Actual Metaphysics of Race

⁸ For Sider's clarification of "Ontologese," see Sider (2011, 171-173).

⁹ For Sider's "revised" definition, see Sider (2011, 49).

Even though Ludwig has provided a valid argument that may be sound as well, it turns out that Ludwig's critique does nothing to undermine the actual metaphysics of race. The latter is partially because Ludwig's critique is not *about* the actual metaphysics of race, it's about a hypothetical metaphysics that he calls 'the new metaphysics of race'.

Remember that the new metaphysics of race is, according to Ludwig (2015, 245), and by definition, constituted by disputes about "one fundamental ontology of race." Furthermore, remember that Ludwig claims that people like Glasgow, Haslanger, Appiah, and Spencer are engaged in one such dispute, the US race debate*, and people like Andreasen, Pigliucci, Kaplan, and Templeton are engaged in another such dispute, the biological race debate*. However, these last two claims are simply false.

For one, the term 'fundamental ontology' is not even a phrase used in actual metaphysics of race. For instance, it does not appear *once* among the actual metaphysics of race that Ludwig (2015, 263-265) cites, and he cites 40 such publications. Second, some actual metaphysicians of race embrace a pluralist ontology for the nature of race in the relevant context. For example, at the beginning of Spencer's (2014, 1026) article on the "national" meaning of 'race' in the US, he concedes that ordinary Americans are using multiple "geographic" and "ethnic" meanings of 'race'. In fact, Spencer (2014, 1026) explicitly says, "Hence, I acknowledge upfront that there are several ways that Americans use 'race'."

However, Ludwig could object here. Specifically, Ludwig (2015, 257) interprets Spencer's focus on the national meaning of 'race' in the US as an endorsement of it being "the only relevant candidate meaning for philosophical debates about the referent of "race" in the United States." While the latter is a possible interpretation of Spencer's project, it's not the most charitable one given how he presents his project at the beginning of his article. Spencer (2014,

1025) begins by saying upfront that his project is merely “to debunk” the idea that “folk racial classification has no biological basis.” Spencer attempts to accomplish that goal by showing that ‘race’, in its national meaning in the current US, is a directly referring term for a biological entity—a set of particular human populations—that presently happens to be biologically real in virtue of being a level of human population structure. Thus, given how Spencer (2014, 1026) presents his own project, his race theory is compatible with there being a pluralist nature of race in the current US context. Furthermore, this interpretation best explains why Spencer (2014, 1026) says that “there are several ways that Americans use ‘race’.”

There are other actual metaphysicians of race who embrace pluralism about the nature of race as well. For instance, Pigliucci and Kaplan (2003, 1162-1163) are happy to grant that both the ecotype and the subspecies are equally legitimate ways of dividing a species into biological races. It’s just that they believe that humans have ecotypes, but not subspecies. In fact, Pigliucci and Kaplan (2003, 1163) explicitly say, “Races, then, can be defined and picked out in a number of ways.”

Finally, there are plenty of actual metaphysicians of race who do not embrace pluralism about the nature of race, but who do entertain pluralism as a metaphysical possibility, which is enough to show that they do not presuppose that there is a single fundamental ontology of race in the relevant context. For instance, after obtaining messy results about how ordinary Americans use ‘race’ and race terms in a widely distributed survey, Glasgow (2009, 75) entertains the possibility that ordinary Americans are sometimes “talking past each other” when they use ‘race’, much like we sometimes do when we use ‘jade’. In fact, Glasgow (2009, 75) explicitly says, “So maybe ‘race’ is used in some contexts to refer to a social kind of thing and in other contexts to a biological kind of thing.” That doesn’t sound like somebody who presupposes that

there is a single fundamental ontology of race in the US context. Now, even though Ludwig's argument is not about actual metaphysics of race, it could still be a relevant critique of actual metaphysics of race. So to that I now turn.

In order to know whether Ludwig's argument succeeds in critiquing the actual metaphysics of race, we need to know more about the debates among actual metaphysicians of race. Clearly, the US race debate* and the biological race debate* are not debates among actual metaphysicians of race. However, the US race debate and the biological race debate are. *The US race debate* is the debate about the nature and reality of race according to what 'race' means in the ordinary discourse of contemporary Americans, but only when 'race' is used to classify humans. The latter debate actually exists because all of the individuals that Ludwig places in the US race debate* have expressed an interest in the focus I've just articulated.¹⁰ *The biological race debate* is the debate about whether humans have any races in a nontrivial biological sense of 'race'. The latter debate actually exists as well.¹¹ These are the two race debates that Ludwig was attempting to critique, and given these distinctions, we can see that Ludwig's argument really isn't relevant to these two debates.

For one, neither the US race debate nor the biological race debate satisfies Hirsch's criterion for a non-substantive dispute. The US race debate is not a merely verbal dispute because racial realists in that debate, such as Haslanger and Spencer, cannot plausibly interpret racial anti-realists in that debate, such as Appiah and Glasgow, as speaking a language in which

¹⁰ For evidence, see Appiah (1996, 42), Glasgow (2009, 15), Haslanger (2012, 133), and Spencer (2014, 1025).

¹¹ For evidence, see Andreasen (1998, 200-201, 205), Pigliucci and Kaplan (2003, 1161-1164), Maglo (2011, 362-363), and Templeton (2013, 262-263).

anti-realist race theories are true, and vice versa. For instance, if Glasgow (2009, 33) is correct about (H1*) being part of the non-negotiable semantic content of ‘race’ in the ordinary discourse of Americans, then Spencer (2014, 1026) is incorrect about ‘race’ directly referring to a set of human populations in the national racial discourse of Americans, and vice versa.¹² The biological race debate is not a merely verbal dispute either. For instance, if Pigliucci and Kaplan (2003, 1165) are correct that humans subdivide into “biologically significant” ecotypes, then Hochman (2013, 347) is incorrect that humans do not subdivide into “meaningful biological units,” and vice versa.

Next, even if the US race debate or the biological race debate is non-substantive in a Ludwagian or Sider-style sense, that fact does not imply a “confusion about metaphysical and normative classificatory issues” as (1) claims. This is because actual metaphysicians of race are adopting a different view of *substantive* metaphysics—namely, one that does not require metaphysical disputes about race to presuppose a single fundamental ontology of race or anything about joint-carving. Thus, while Ludwig’s argument is relevant to the hypothetical new metaphysics of race, it doesn’t make contact with actual metaphysics of race.

Interestingly, when Ludwig defines ‘the new metaphysics of race’, he anticipates the worry that his focus on it may mischaracterize actual metaphysics of race. In response, Ludwig (2015, 245) says, “However, I do not want to engage in a verbal dispute about the meaning of ‘metaphysics of race’... this article only challenges a certain type of metaphysics of race while proposing an alternative deflationist and normative metaphysics of race.” However, this reply is

¹² (H1*) is the claim that a race is, at least, a group of human beings that is distinguished from other groups of human beings by visible physical features, of the relevant kind, that the group has to some significantly disproportionate extent (Glasgow 2009, 33).

perplexing because if the new metaphysics of race is a purely hypothetical metaphysics that does not describe the disputes in actual metaphysics of race (as I've shown), and, in addition, if the disputes in actual metaphysics of race already do away with monist and fundamentalist assumptions about race (as I've shown), it's hard to imagine what the purpose is for lodging Ludwig's critique in the first place. In any case, we can rest assured that actual metaphysicians of race are immune to Ludwig's critique because they've already been vaccinated against monist and fundamentalist assumptions about race.

5. Closing Remarks

In this paper, I've shown that Ludwig's critique of the new metaphysics of race is irrelevant to the actual metaphysics of race. However, I've said little about the conditions of substantivity that actual metaphysicians of race adopt. In addition to the bare minimum of "not talking past one another" (Glasgow 2009, 28), actual metaphysicians of race embrace disputes about how certain linguistic communities actually use 'race' (e.g. Pigliucci and Kaplan 2003, 1162-1163; Glasgow 2009, 6), and embrace disputes about how certain linguistic communities should use 'race' (e.g. Haslanger 2012, 221-247; Hochman 2014, 80). However, actual metaphysicians of race do not embrace disputes that have unimportant social and scientific consequences. For instance, Haslanger (2012, 300) motivates the US race debate by pointing out that engaging in it will help us frame and evaluate social policies and appropriately address stubborn inequalities in health. Also, Pigliucci and Kaplan (2003, 1170) point out that engaging in the biological race debate can help biologists debunk hereditarian hypotheses about race and intelligence, yield insights into human evolutionary history, and yield insights into human migration history.

Interestingly, the criteria for substantive metaphysics that actual metaphysicians of race adopt make the metaphysical disputes in the actual metaphysics of race more akin to metaphysical disputes in the philosophy of science (e.g. the species debate, the nature of natural kinds, the ontic structural realism debate, etc.) than those in mainstream analytic metaphysics (e.g. debates about the nature of fundamentality, grounding, modality, substantivity, etc.). For instance, Matthew Slater's (2015) stable property cluster theory of natural kinds has a real shot at explaining why some kinds support epistemically reliable inductions in a domain while others don't, which could help systematic biologists achieve more agreement about how they should classify organisms into species and higher taxa. So, much like disputes in the actual metaphysics of race, there are practical payoffs to science or society for engaging in metaphysical disputes in the philosophy of science. However, mainstream analytic metaphysics does not guarantee a payoff for science or society. For instance, what exactly is the payoff for science or society in debating about "the" nature of substantive metaphysics?

Perhaps Sider (2011, 47) sums up my point best when he says, "... this concept is not intended to apply to everything that might justly be called "nonsubstantive". For example, it isn't meant to apply to equivocations between distinct lexical meanings (as in a dispute over whether geese live by "the bank", in which one disputant means river bank and the other means financial bank)... Nor is it meant to capture the shallowness of inquiry into whether the number of electrons in the entire universe is even or odd (an inquiry that is substantive in my sense, but pointless)."

References

Andreasen, R. O. (1998). A New Perspective on the Race Debate. *The British Journal for the Philosophy of Science*, 49(2), 199-225.

- Appiah, K. A. (1996). Race, Culture, Identity, Misunderstood Connections. In K. A. Gutmann, *Color Conscious* (pp. 30-105). Princeton: Princeton University Press.
- Glasgow, J. (2009). *A Theory of Race*. New York: Routledge.
- Glasgow, J., Shulman, J., & Covarrubias, E. (2009). The Ordinary Conception of Race in the United States and Its Relation to Racial Attitudes: A New Approach. *Journal of Cognition and Culture*, 9, 15-38.
- Haslanger, S. (2012). *Resisting Reality*. Oxford: Oxford University Press.
- Hirsch, E. (2005). Physical-Object Ontology, Verbal Disputes, and Common Sense. *Philosophy and Phenomenological Research*, 70(1), 67-97.
- Hochman, A. (2013). Against the New Racial Naturalism. *The Journal of Philosophy*, CX(6), 331-351.
- Hochman, A. (2014). Unnaturalised racial naturalism. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 46, 79-87.
- Kaplan, J., & Winther, R. (2013). Prisoners of Abstraction? The Theory and Measure of Genetic Variation, and the Very Concept of "Race". *Biological Theory*, 7(1), 401-412.
- Ludwig, D. (2015). Against the New Metaphysics of Race. *Philosophy of Science*, 82(2), 244-265.
- Maglo, K. (2011). The Case against Biological Realism about Race: From Darwin to the Post-Genomic Era. *Perspectives on Science*, 19(4), 361-390.
- Pigliucci, M., & Kaplan, J. (2003). On the Concept of Biological Race and Its Applicability to Humans. *Philosophy of Science*, 70(5), 1161-1172.
- Sider, T. (2011). *Writing the Book of the World*. Oxford: Oxford University Press.
- Slater, M. (2015). Natural Kindness. *The British Journal for the Philosophy of Science*, 66(2), 375-411.
- Spencer, Q. (2014). A Radical Solution to the Race Problem. *Philosophy of Science*, 81(5), 1025-1038.
- Templeton, A. R. (2013). Biological Races in Humans. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 44(3), 262-271.
- Zack, N. (2002). *Philosophy of Science and Race*. New York: Routledge.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

Measurement of Statistical Evidence: Picking Up Where Hacking (et al.) Left Off

Abstract Hacking's (1965) Law of Likelihood says – paraphrasing– that data support hypothesis H_1 over hypothesis H_2 whenever the likelihood ratio (LR) for H_1 over H_2 exceeds 1. But Hacking (1972) noted a seemingly fatal flaw in the LR itself: it cannot be interpreted as the degree of “evidential significance” across applications. I agree with Hacking about the problem, but I don't believe the condition is incurable. I argue here that the LR *can* be properly calibrated with respect to the underlying evidence, and I sketch the rudiments of a methodology for so doing.

Introduction

The “likelihoodist,” or “evidentialist,” school of thought in statistics is well known among philosophers, more so perhaps than among scientists or even statisticians, in large part due to Hacking (1965). One way to distinguish evidentialism from the other major schools – frequentism and Bayesianism – is to note that evidentialism alone focuses on the assessment of statistical evidence as its principal task, rather than decision-making or the rank-ordering of beliefs.¹

¹ Hacking himself generally prefers the term “support” over “evidence,” as does Edwards (1992), but other representatives of this school (Good 1950; Barnard 1949; Royall 1997) refer to an equivalent concept as “evidence.” I prefer “evidence,” since this is the familiar, albeit vague, word for what we are trying to illuminate; and I prefer “evidentialist” over “likelihoodist” as the name of the school, since the former highlights a key distinction

Veronica J. Vieiland
Philosophy of Science Assoc Biennial Meeting 2016

It might be thought, therefore, that evidentialism would be the predominant approach to statistical inference in science, where quantifying evidence is usually the main objective. (If you don't agree, try getting scientists to stop using the p-value as a measure of the strength of the evidence!) But frequentism, and to a lesser extent Bayesianism, predominate in the scientific literature, while evidentialism is virtually unseen. Why is this? I'm going to argue here that the fault lies with evidentialism's failure thus far to address the problem of calibrating the units in which evidence is to be measured. Since meaningful calibration is the *sine qua non* of scientific measurement, this turns out to be the loose thread that causes the cloth to unravel when we pull on it.

Before proceeding it may be worth noting some things I will and will not be talking about. First, I am concerned only with *statistical* evidence, and will not be considering the concept of evidence as it appears in other contexts, e.g., in legal proceedings. Second, I will treat statistical evidence as a *relationship* between data and hypotheses under a model that can be expressed in the form of a likelihood (as defined below). On this view, data do not possess inherent evidential meaning on their own, but only take on meaning in the context of their relationships to particular hypotheses, with the nature of those relationships governed by the form of the likelihood. I will not be concerned here with measurement problems associated

between this school and the others. By contrast, likelihood features prominently in all modern statistical frameworks.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

with the data themselves.² Third, I am interested here solely in addressing the question of whether this relationship between data and hypotheses can be rigorously quantified. If the answer is yes, then presumably the degree of evidence could play a role in decision making (deciding how strong is strong enough when it comes to evidence) or in guiding belief, but I will not be addressing these topics here. It is one hallmark of evidentialist reasoning that statistical evidence is treated independently of these matters.

The remainder of the paper is organized as follows. In section (1) I articulate the central evidence calibration problem (ECP), and suggest reframing it in measurement terms. In section (2), I consider ways in which evidentialism's preoccupation with so-called "simple" hypotheses (as defined below) has constricted the theory, masking the true nature of the underlying measurement problem, and also obscuring the solution. In section (3) I illustrate a methodology for beginning to address the ECP once the restriction to simple hypotheses is relaxed. In section (4) I briefly consider what changes would be required to axiomatic foundations in order to accommodate this methodology while remaining true to the spirit of evidentialism's original motivating arguments.

(1) The Evidence Calibration Problem (ECP)

At the heart of evidentialism is Hacking's (1965) familiar Law of Likelihood, which says in essence that data support one statistical hypothesis H_1 over another hypothesis H_2

² In common usage "evidence" is often used to refer to what I am calling *data*, but "evidence" also has this other sense of being a *relationship* between data and hypotheses. In order to maintain this distinction, I will call the data "data" and the relationship "evidence."

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

whenever the likelihood ratio (LR) for H_1 over H_2 exceeds 1. But Hacking (1972) pointed out a problem in assigning any particular interpretation to the magnitude of the LR. In his review of Edwards (1992, orig. 1972), he says:

“Now suppose the actual log-likelihood ratio between the two hypotheses is r , and suppose this is also the ratio between two other hypotheses, in a quite different model, with some evidence altogether unrelated to [the original data]. I know of no compelling argument that the ratio r ‘means the same’ in these two contexts.”³ (p. 136)

Thus we can say that, for one experiment, data support hypothesis H_1 over hypothesis H_2 with $LR = 2$, and, for another experiment, that a different set of data support H_3 over H_4 with $LR = 20$; but we cannot say anything definite about how much more the second set of data supports H_3 over H_4 relative to the amount by which the first set supports H_1 over H_2 .

Edwards was well aware of this problem, saying expressly that “we shall not be attempting to make an absolute comparison of *different* hypotheses on *different* data.” (p. 10). But

Hacking’s point cuts deep. *If the numerical value of the LR cannot be meaningfully compared across applications, in what sense is it meaningful in any one application?*

³ Here Hacking is using “evidence” in the sense of what I am calling *data*; however, he goes on to describe what he has in mind in terms of levels of “evidential significance.” He refers to the *log* LR as this is the form preferred by Edwards. Note that Hacking already appears to have been alluding to this problem in Hacking (1965), vide p. 61.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

Hacking's criticism points to a fundamental problem for evidentialists, who appear to be able to say *whether* given data support H_1 over H_2 , but not by *how much* they support H_1 .⁴ This is on the face of it metaphysically perplexing, but also, it leaves a gap between *support*, as Hacking's Law defines it, and a truly quantitative *weight of evidence*, which would be far more useful scientifically if only we could work out how to evaluate it.

Following the core arguments in Barnard (1949), Hacking (1965) and Edwards (1992), I will assume that the LR is the key quantity in any cogent theory of statistical evidence. But the Law of Likelihood is more specific than this assumption: it assigns a particular importance to one very narrowly conceived *aspect* of the LR, a fact that is obscured by evidentialism's focus on simple hypotheses, to which I turn next.

Before doing so, I note that resolving Hacking's problem requires unpacking his phrase 'means the same'. I think that this must be understood as 'means the same with respect to the underlying evidence,' a locution that lands us solidly in *measurement* territory. We must be able to think in terms of the underlying evidence, as something we can – at least in the abstract – conceive of independently of how we measure it. The question then becomes: How do we establish meaningful measurement units for evidence, so that a given measurement value always 'means the same' *with respect to the evidence*? This is the ECP.

And here, in a nutshell, is the evidentialist's difficulty in addressing the ECP. The LR for a simple hypothesis comparison (see below) is a single number, thus, the evidentialist is lured

⁴ Royall (1997) is the only one as far as I know who argues that the magnitude of the LR *does* express strength of evidence in a comparable manner across applications. But I think his arguments on this point fail for reasons articulated in Forster & Sober (2004).

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

into the claim that “the LR *is* the evidence.” To see the danger here, consider a mercury thermometer reading 80°F. We might say, “the temperature is 80°,” but this is a circumlocution for “80 is the numerical value we assign, on the Fahrenheit scale, to the underlying temperature.” Now suppose that rather than degrees, only units of volume V are annotated on the sides of the glass. We might be tempted to say “ V is the temperature,” but now this statement is not merely a circumlocution, it is also an error. V alone does not tell us the temperature; we must, at the least, also take into account the pressure. To insist that temperature can be represented by volume alone, or by pressure alone, or by any other single thing that can be readily and directly measured, is to mistake the nature of temperature. Just so, I am going to argue that *the simple LR mistakes the nature of evidence*, by obscuring the fact that the evidence itself is not a number, and moreover, that the evidence is not any single thing that can be readily and directly measured, but instead, it is a function of (at least) two measurable things.

(2) The Insidiousness of Simple Hypotheses

To begin with, we need to define *likelihood*:

“The likelihood, $L(H|R)$, of the hypothesis H given data R , and a specific model, is proportional to $P(R|H)$, the constant of proportionality being arbitrary.” Edwards (1992)

(p. 9)

Two key points are familiar: (i) likelihood represents a feature of an hypothesis given data, not the other way around; and (ii) likelihood is related to but not the same as probability,

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

since it is defined only up to an arbitrary multiplicative factor and therefore does not follow the Kolmogorov axioms. I will not rehearse the advantages of likelihood in spelling out a theory of statistical evidence, but suffice it to say that likelihood enables inferences to proceed independently of what are, arguably, extraneous features of study design, including the sampling distribution of all those observations that might have occurred but didn't.

There is a third important feature of this definition as well, and this regards the nature of the *hypotheses* to which the definition is intended to apply. Edwards is, as always, explicit:

“An essential feature of a statistical hypothesis is that its consequences may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached.” (p. 4)

This precludes consideration of likelihoods involving *composite* hypotheses. For instance, in the context of a coin-tossing experiment in which x independent tosses have landed heads and y have landed tails, and letting $\theta = P(\text{heads})$, one can write the likelihood $L(\theta=0.1|x, y)$, or $L(\theta=0.2|x, y)$. These likelihoods involve “simple” hypotheses, in which θ is assigned a single numerical value, so that the corresponding probability $P(x, y|\theta)$ returns a single number on the probability scale for each possible outcome (x, y) . But one can *not* write $L(\theta=0.1$ or $\theta=0.2|x, y)$, because the latter involves a “composite” hypothesis, which does not assign a definite probability to the observed outcome. To know the probability of observing (x, y) under the hypothesis “ $\theta=0.1$ or $\theta=0.2$,” we would need not only to know the probability of (x, y) for each θ , but also, we would need to know the prior probabilities of $\theta=0.1$ and $\theta=0.2$. As these prior probabilities lies outside the likelihood, they are not admissible on the

Veronica J. Vieldand
Philosophy of Science Assoc Biennial Meeting 2016

evidentialist view.

But even the simplest examples of statistical reasoning generally involve hypotheses that appear on the face of things to be composite; e.g., we might be interested in whether the coin is biased toward tails or fair, which would appear to involve the improperly formed hypothesis $\theta < 0.5$. This situation is handled by treating composite hypotheses “solely on the merits of their component parts” (Edwards, p. 5). Thus in forming the LR corresponding to ‘coin is biased toward tails’ vs. ‘coin is fair,’ we would need to consider separately the (infinitely many) simple LR’s in the form $L(\theta = \theta_i | x, y) / L(\theta = 0.5 | x, y)$, for each possible i^{th} value of $\theta \leq 0.5$. Now the LR is a function of θ , not a single number (Figure 1).

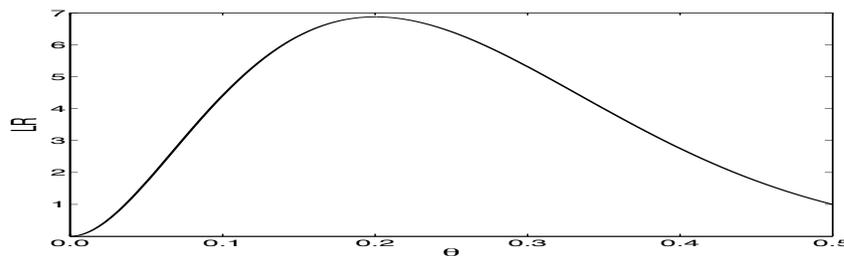


Figure 1 LR as a function of θ for $x = 2, y = 8$.

In practice it seems that what is important is not so much the proscription against composite hypotheses, but rather the prescription for how they may be interpreted. We can graph the LR as a function of θ , as if we were admitting composite hypotheses, but we can only make statements like “ $\theta = 0.2$ is supported over $\theta = 0.5$, on given data, by $LR = 6.9$,” while

Veronica J. Vieiland
Philosophy of Science Assoc Biennial Meeting 2016

“ $\theta=0.1$ is supported over $\theta=0.5$, on those same data, by $LR=4.4$.”⁵ But as a practical matter, the graph is not a sufficiently concise summary for general scientific applications. We still need some way to reduce the function $LR(\theta)$ to a single number summarizing the strength of the evidence.

And this is where we get into trouble, because focus shifts naturally to the *maximum* LR (MLR), which occurs over the best supported value – the maximum likelihood estimate (m.l.e.) – of θ . Indeed, given that we are only allowed to make statements about one simple hypothesis comparison at a time, the MLR, itself a ratio of two simple likelihoods, appears as the best single constituent LR to use as a summary feature of the LR graph. (Below I consider how relaxing the requirement that hypotheses must be simple frees us up to consider other features.) We have now successfully summarized the *function* $LR(\theta)$ as a single number, the MLR, but this summary is tethered to the m.l.e.. We appear to have answered the question: How well supported is the m.l.e. compared to (one or more individual) alternative values of θ ? But that is not the question we asked initially, which was about the evidence.⁶

The m.l.e. of θ arrives on the scene as a seemingly innocuous point of special interest, the value that corresponds to the maximum support, but it rapidly takes over, embroiling us in a downward spiral of increasingly perplexing difficulties. One immediate issue with relying on the MLR to summarize the evidence (continuing to focus for ease of discussion on the coin-

⁵ Moreover we can only make such statements when both the data and the form of the likelihood are the same in the numerator and the denominator of the LR, for only in such cases will the constants of proportionality cancel.

⁶ Hacking (p. 28 ff.) makes clear the conceptual reasons for keeping estimation and evidence (or support) separate.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

tossing example, in which maximization occurs only in the numerator of the LR), is that $MLR \geq 1$: the MLR can only show evidence in favor of the numerator but never in favor of the denominator. This is problematic, like using a thermometer in which the mercury is prevented from receding.

Another problem with the MLR is that it begs the question of measurement scale in a particularly obvious way, because its evidential meaning would appear to require some kind of adjustment to compensate for the maximization itself. The more parameters we maximize over (again, for ease of discussion, assuming maximization occurs only in the numerator), the larger the MLR becomes. How are we to separate the portion of the MLR reflecting the evidence from the portion representing an artifact of the process of maximization? It becomes particularly hard to retain the fiction that the numerical value of the *maximum* LR has some prima facie meaning with respect to the underlying evidence, regardless of the number of parameters over which the LR is maximized.

There is a third, more subtle but at least as damaging, difficulty with summarizing evidence via MLRs. Simple LRs can be multiplied across two data sets, but MLRs can not be multiplied. Rather, to obtain the MLR based on two sets of data, we first combine the data to find the new m.l.e., which is a kind of weighted average of the two original m.l.e.s, and then we find the new MLR with respect to this average m.l.e. on the combined data. Now consider a situation in which data set D_1 favors H_2 by some substantial amount, and D_2 also favors H_2 , but by a lesser amount. In such situations it is not uncommon for the combined support for H_2 to be less than the original support on D_1 alone. But this is not how *evidence* behaves:

Veronica J. Wieland
Philosophy of Science Assoc Biennial Meeting 2016

strong evidence for H_2 followed by weaker evidence also supporting H_2 ought to lead to *stronger* evidence for H_2 , not intermediate evidence. (A blood type match following a DNA match does not lessen the evidence that the defendant was at the crime scene.⁷) This means that we cannot in practice differentiate between situations in which new data are truly diminishing the evidence, and situations in which the evidence is in fact increasing but the MLR at the average m.l.e. goes down anyway. This tendency of the MLR to “average” across combined data is entirely due to its dependence on the m.l.e.; simple LRs do not share this defect.⁸

Of course none of this need surprise unreconstructed evidentialists, who, after all, disavowed composite hypotheses – and therefore any need for maximization – from the start. But then beyond the simplest of examples, we are left with an irreducible graph of the component simple LRs, not a single number. This is true already in single-parameter cases; the problem is only exacerbated in higher dimensions.

There is also the matter of masking the nature of the real problem: by focusing initially only on those situations in which the LR is a single number, we missed Hacking’s *measurement* question, how do we ensure that this number always ‘means the same’? It is only when we consider composite hypotheses that it becomes clear we were never warranted

⁷ This example was suggested by Hasok Chang.

⁸ This issue plays a salient role in the current “crisis” of non-replication of statistical findings in the biomedical and social sciences, where the tendency of p-values and MLRs to “regress to the mean” upon attempts to replicate initial findings is widely interpreted as meaning that the evidence has gone down. In the absence of a properly behaved evidence measure, however, this conclusion is entirely unwarranted.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

in the first place in assuming that the face value of the LR for a simple vs. simple hypothesis comparison *is* the evidence. Composite hypotheses force us to think in terms of the LR graph, which, precisely because it is not a single number, immediately raises the issue of which *feature(s)* of the graph might be relevant to the evidence. Composite hypotheses are crucial, not only because they are scientifically relevant, but also, because they beg a question all but hidden as long as we focus only on simple hypotheses.

The urge to sidestep the problem of the evidential interpretation of the MLR is the reason evidentialists have been reluctant to admit composite hypotheses into their formalism in the first place. But it is fair to say that they have failed to provide any viable alternative to the MLR as the summary measure of evidence strength in practice. The preoccupation with simple hypotheses has entailed inherent difficulties for the program, and it has also masked a basic underlying calibration issue. The good news, I believe, is that it has also been masking the possibility of a solution.

(3) Towards a Solution to the Measurement Calibration Problem

Consider again the coin-tossing experiment and $LR(\theta)$ as shown in Figure 1. Let us suppose, following the spirit if not the letter of the Law of Likelihood, that all of the evidential information is captured, somehow, in this graph. What *feature(s)* of the graph should we take as representing the degree of evidence?

The MLR of course is one possibility, but I have already stated some objections to this option. An alternative would be to use the *area* under the graph (ALR). (Note that this is

Veronica J. Vieiland
Philosophy of Science Assoc Biennial Meeting 2016

only possible if we allow ourselves to consider the truly composite hypothesis $\theta < 0.5$, because the ALR requires simultaneous consideration of all of the constituent simple hypotheses.⁹) But while we're at it, why not also consider using *sets of features* of the graph? For instance, the evidence might be a function of both the MLR and the ALR, e.g., their product, or their ratio. What we need is a methodology for figuring out which among the many possibilities is the correct one.

The methodology I propose is quite simple, at least to begin with. Let's consider the *behavior of candidate evidence measures* in situations where we have clear intuitions regarding the *behavior of evidence*, and see which of our candidate measures behaves like the object of measurement, the evidence. Here I will illustrate using coin-tossing "thought experiments" to discover patterns of behavior of the evidence with changes in data, considering the evidence that the coin is either biased toward tails or fair. I propose that, perhaps with a little persuasion, I could convince you that the following patterns capture *what we mean* when we talk about statistical evidence in this context. (Here I summarize the data in terms of n =the number of tosses, and x/n =the proportion of tosses that land heads.)

- (i) Evidence as a function of changes in n for fixed x/n For any given value of x/n , the evidence increases as n increases. The evidence may favor bias (e.g., if $x/n = 0.05$) or no bias (e.g., if $x/n = 1/2$), but in either case it gets stronger with increasing n .

⁹ The ALR is proportional in this simple example to the Bayes factor under a uniform prior on θ , which is sometimes interpreted in Bayesian circles as a measure of evidence strength; it is also proportional to the relative belief (Evans 2015), another Bayesian proposal for measuring evidence. But the ALR itself does not involve a prior, so I see no *prima facie* reason for the evidentialist to balk at this suggestion, once composite hypotheses are allowed.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

(ii) Evidence as a function of changes in x/n for fixed n If we hold n constant but allow x/n to increase from 0 up to, say, 0.20, the evidence favoring ‘coin is biased’ diminishes: i.e., the evidence for bias is stronger the further x/n is from $\frac{1}{2}$. But we have also already noted that when x/n is close to $\frac{1}{2}$ the evidence favors ‘coin is fair.’ Therefore, as x/n continues to approach $\frac{1}{2}$, at some point the evidence will shift to favoring ‘coin is fair,’ and from that point, the evidence for ‘coin is fair’ will increase the closer x/n is to $\frac{1}{2}$.

(iii) Rate of evidence change as a function of changes in n for fixed x/n For given x/n , as n increases the evidence *increases more slowly* with fixed increments of data. E.g., consider evidence in favor of bias with one additional tail (T), following T, or TT, or TTT. When the number of tails in a row is small (i.e., when there is weak evidence favoring bias), each subsequent T makes us that much more suspicious that the coin is biased. But suppose we have already observed 100 Ts in a row: now one additional T changes our sense of the evidence hardly at all, as we are already quite positive that the coin is not fair.¹⁰

(iv) x/n as a function of changes in n (or vice versa) for fixed evidence It follows from (i) and (ii) that in order for the *evidence* to remain constant, n and x/n must adjust to one another in a compensatory manner. E.g., if x/n increases from 0 to 0.05, in order for the evidence to remain the same n must increase to compensate; otherwise, the evidence would go down, following (ii) above. By the same token, it is readily verified that if (i)

¹⁰ This underscores the point made above that evidence is not inherent in the data (say, a single toss T), but rather, evidence is a relationship between the data and the hypotheses that depends on context.

Veronica J. Vieland
Philosophy of Science Assoc. Biennial Meeting 2016

and (ii) hold, then as x/n continues to increase, at some point n must begin to decrease in order to hold the evidence constant as the evidence shifts to favoring ‘coin is fair.’ Note that at this point we have not mentioned probability distributions, likelihoods, or parameterization of the hypotheses. These patterns characterize evidence in only a very informal, vague manner. However, by the same token, they exhibit a kind of generality: they derive from our general sense of evidence, from what we *mean* by statistical evidence before we attempt a formal mathematical treatment of the concept.

Can we find a precise mathematical expression that exhibits these patterns? As illustrated in Figure 2, the ratio $RLR = MLR/ALR$ exhibits all of the expected behaviors. By contrast, neither MLR nor ALR shows all four of these patterns. For instance, MLR, as already noted, cannot show increasing evidence in favor of H_2 because it can never favor H_2 in the first place; and both MLR and ALR increase exponentially in n for fixed x/n rather than showing the concave-down pattern in 2(a).

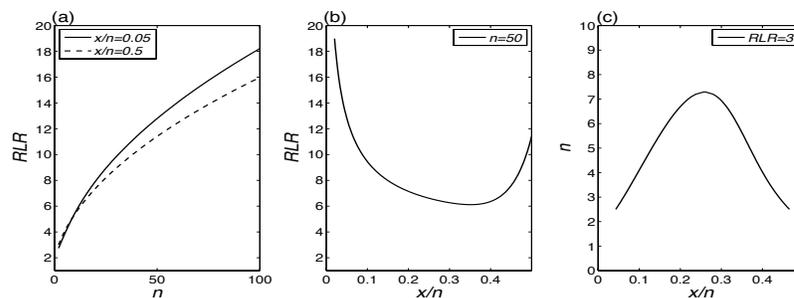


Figure 2 Patterns of behavior of RLR for coin-tossing thought experiments: (a) Patterns (i) and (iii); (b) Pattern (ii); (c) Pattern (iv).

Veronica J. Wieland
Philosophy of Science Assoc Biennial Meeting 2016

Of course none of this proves that RLR is the correct, or optimal (or properly calibrated) measure of evidence. But this style of reasoning buys us an important methodological tool. Whichever features of the LR graph we consider and however we combine them, we must be able to show that the resulting evidence measure *behaves like the evidence*. When proposing candidate evidence measures anything goes, but only those candidates that behave appropriately remain on the ballot. And even in this very simple example, two obvious candidates – the MLR and the ALR – have already dropped out of contention.

Of course, there is no reason to assume that what works in this simple case (RLR) will work in more complicated cases, nor have we yet resolved the ECP's fundamental calibration issue. Establishing that a measure behaves like the object of measurement is only a first step, but it is a vital step not previously taken. It provides an "empirical" measurement scale, not an absolute scale, much as early thermoscopes provided good experimental tools while falling short of proper, absolute, calibration (Chang 2004).¹¹ Projecting an empirical measure onto an absolute scale requires a broader theoretical foundation, but one needs the empirical measure first. My point here is simply that confronting the ECP head on, and in the context of composite hypotheses, opens the door for the first time to the possibility of establishing a proper measurement scale for statistical evidence.

Note too that the coin-tossing exercise suggests the existence of an *equation of state* involving the three quantities (n , x/n and the evidence), such that fixing any one quantity

¹¹ Indeed, the ECP poses what Chang calls a "nomic" measurement problem, much like the nomic problem of temperature measurement. What I am describing here is a necessary but not sufficient stage in resolving a nomic problem.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

while allowing a second one to change requires a specific compensatory change in the third. This in turn suggests a new, and potentially very powerful, way to think about the laws governing the behavior of LRs. I'm not aware of any evidentialist work that considers such equations, but I see no reason that an evidentialist-at-heart should be prohibited from pursuing their study.

(4) Relaxing the Foundations To Include Composite Hypotheses

In order to tackle the ECP in the terms of the preceding section, we need to amend the foundations of evidentialism, but only slightly. I propose the following changes. First, let's retain Edwards' definition of likelihood, as quoted above, but insert the word "simple" (which is tacit in Edwards' original statement): "The likelihood, $L(H|R)$, of a *simple* hypothesis H given data R , and a specific model, is proportional to $P(R|H)$, the constant of proportionality being arbitrary." Second, we can again add the word "simple" to his characterization of a statistical hypothesis: "An essential feature of a *simple* statistical hypothesis is that its consequences may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached." But we can now add a definition of likelihood for a composite hypothesis: "A *composite* hypothesis H given data R , and a specific model, is the set of all constituent simple hypotheses, defined up to a single constant of proportionality." Thus the essential feature of a *composite* hypothesis is that *each of its constituent simple hypotheses* may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached. We can now use this definition

Veronica J. Vialand
Philosophy of Science Assoc Biennial Meeting 2016

of a composite hypothesis to define the corresponding composite likelihood, as the set of all constituent simple likelihoods.

Under my proposal, the spirit of the Law of Likelihood can be retained: We can say that all of the *evidential information* conveyed by given data regarding a comparison between two hypotheses on a particular model is contained in the LR, where, under the expanded definition of hypotheses, the LR is understood to be a function of all unknown parameters, or better still perhaps, a *graph*. This can equivalently be read as a definition of *evidential information*, as whatever changes the LR graph.¹² But the idea that the (simple) LR itself expresses the degree or weight of the evidence must be abandoned. What I have attempted to argue here is that there is at least the possibility of replacing this notion with something more useful.

Discussion

Evidence is a general and vague term in science. Statistical evidence is a narrower concept, but it still inherits some of this vagueness. One way to tackle a general and vague term is by seeking a precise definition that maintains full generality, but of course, this might not be possible. Weyl (1952) has suggested another approach:

“To a certain degree this scheme is typical for all theoretic knowledge: We begin with some general but vague principle, then find an important case where we can give that

¹² I borrow this idea from Frank (2014), who defines *information* as whatever changes a probability distribution.

Veronica J. Vieiland
Philosophy of Science Assoc Biennial Meeting 2016

notion a concrete precise meaning, and from that case we gradually rise again to generality... and if we are lucky we end up with an idea no less universal than the one from which we started. Gone may be much of its emotional appeal, but it has the same or even greater unifying power in the realm of thought and is exact instead of vague.” (p. 6)

Can evidentialism be redeemed and made truly useful to science? Of course I have not proved that the answer is yes. But in section (3) I illustrated a case in which we appear to be able to give the vague concept of statistical evidence a concrete, precise meaning, via the quantity $RLR=MLR/ALR$. It remains to be seen whether it is possible to rise again to generality from this first step. But for those of us who agree with most of what Barnard, Hacking and Edwards have to say on the subject, it seems worthwhile to see how far we can take this line of reasoning. This also seems to be a singular opportunity for philosophers of science to step into the breach and at least *try* to solve a problem that has long stood between one of the needs of science – for well-behaved quantitative measures of evidence – and the capabilities of conventional statistical methodologies.

References

- Barnard G.A. "Statistical Inference." *J Royal Stat Soc* XI, no. 2 (1949):115-39.
- Chang H. *Inventing Temperature: Measurement and Scientific Progress*. New York:Oxford UP, 2004.
- Edwards A.W.F. *Likelihood*. Baltimore:Johns Hopkins UP, 1992. Orig. Cambridge UP, 1972.
- Evans M. *Measuring Statistical Evidence Using Relative Belief*, Monographs on Statistics and Applied Probability. Boca Raton:CRC Press, Taylor & Francis Group, 2015.
- Forster M, Sober E. "Why Likelihood?" In *The Nature of Scientific Evidence*, Taper & Lele eds., 153-90. Chicago:Chicago UP, 2004.
- Frank S.A. "How to Read Probability Distributions as Statements About Process." *Entropy* 16(2014):6059-98.
- Good I. J. *Probability and Weighing of Evidence*. London:Griffon, 1950.

Veronica J. Vialand
Philosophy of Science Assoc Biennial Meeting 2016

Hacking I. *Logic of Statistical Inference*. London:Cambridge UP, 1965.

———. "Review of Edwards' Likelihood." *British J Phil of Sci* 23(1972): 132-37.

Royall R. *Statistical Evidence: A Likelihood Paradigm*. London:Chapman & Hall, 1997.

Weyl, Hermann. *Symmetry*. Princeton UP, 1952.

What Basic Emotions Really Are

Encapsulated or Integrated?

Abstract: While there is ongoing debate about the existence of basic emotions (BEs) and about their status as natural kinds, these debates usually carry on under the assumption that BEs are encapsulated from cognition and that this is one of the criteria that separates the products of evolution from the products of culture and experience. I aim to show that this assumption is entirely unwarranted, that there is empirical evidence against it, and that evolutionary theory itself should not lead us to expect that cognitive encapsulation marks the distinction between basic and higher cognitive emotions. Finally, I draw out the implications of these claims for debates about the existence of basic emotions in humans.

1. Introduction

It is widely held among emotion theorists that there is some theoretically interesting distinction between basic and higher cognitive emotions. On this picture, basic emotions (BEs) are primarily structured by evolution whereas higher cognitive emotions are substantially structured by either culture or individual experience. While there is ongoing debate about the existence of BEs and about their status as natural kinds, these debates usually carry on under the assumption that BEs are encapsulated from cognition and that encapsulation is one of the criteria that separates the products of evolution from the products of culture and experience. I aim to show that this assumption is entirely unwarranted, that there is empirical evidence against it, and that evolutionary theory itself should not lead us to

Isaac Wiegman
10/19/2016

expect that cognitive encapsulation marks the distinction between basic and higher cognitive emotions. Finally, I draw out the implications of these claims for the existence of basic emotions in humans.

In the following section, I characterize the received view of BEs, which holds (among other things) that BEs are solutions to *basic life problems* in our evolutionary past. Then I consider and reject some of the reasons to think that BEs are cognitively encapsulated. In the second section, I provide an example of a BE in rodents that bears the marks of cognitive integration (as opposed to encapsulation). The basic life problem that likely shaped this emotion appears to demand substantial cognitive integration. In the third section, I draw out the implications for a current debate in emotion theory concerning the existence of BEs in humans.

2. Basic Emotions

BEs – including anger, fear, happiness, sadness, disgust, and surprise (for an extended list, see Ekman & Cordaro, 2011) – are thought to be human-typical behavioral syndromes that include involuntary facial expressions of emotion, physiological changes (e.g. in heart rate, blood pressure, and hormone levels), and changes in bodily posture (including bodily social displays and orienting responses). According to BE theory, these syndromes have a similar kind of evolutionary explanation and similar neural and psychological mechanisms. Specifically, they each evolved to address basic life problems or adaptive problems (such as

Isaac Wiegman
10/19/2016

resource competition, avoidance of predators and avoidance of poisons and parasites). Some of these basic life problems are ones that we share with non-human animals.

Moreover, the elicitation and production of these syndromes (including the coordination of various response components) are supposed to be explained by *automatic appraisal mechanisms* and *affect programs*, respectively (Ekman, 1977, 1999). For instance, affect programs explain phenomena observed in experiments that ask people to distinguish photographs of facial expressions of emotions, connect these expressions with emotion terms, or rate their appropriateness in response to vignettes (for an overview, see Ekman, 2003). They are also supposed to explain the results of experiments that connect facial expressions with changes in physiological response components (Ekman, Levenson, & Friesen, 1983; Levenson, Ekman, & Friesen, 1990). To generalize, affect programs are introduced to explain the observed coordination of various response components and the cross-cultural production of these various syndromes (which is thought to explain widespread recognition of facial expressions across cultures).

3. Unwarranted Assumptions Concerning Cognitive Integration

Many emotion theorists claim that BEs lack cognitive integration. In this section, I argue that these claims are based on unwarranted assumptions.

Assumption 1: Cognitively Integrated only if Informationally Integrated

In most cases, questions about the integration of emotions with cognition concern the possibility that emotions are modular in Fodor's (1983) sense. This depends (among other

Isaac Wiegman
10/19/2016

things) on whether they can store *information* that cognitive systems cannot access (*informational encapsulation*); or whether *information* from other cognitive systems can interfere with the operations of an emotion (*cognitive penetrability*); or whether people have conscious access to emotional processes or merely their outputs (*opacity*); or whether the *information* that an emotion provides is general as opposed to specific (which would imply *shallow outputs*). These are some of the more well-known marks of cognitive integration or its absence, encapsulation.

Philosophers and psychologists alike usually proceed under the assumption that integration with cognition depends entirely on whether information is integrated in these ways. These assumptions translate to discussions about BEs, where evidence for lack of *informational* integration is sometimes used as evidence for lack of *cognitive* integration *simpliciter*:

Three other types of evidence suggest that [basic] emotion processes can operate independently of cognition. Emotions have been induced by unanticipated pain..., manipulation of facial expressions..., and changing the temperature of cerebral blood... In all these conditions the immediate cause of the emotion was noncognitive. (Izard, 1992, p. 563, see also his 2007)

Here, Izard apparently assumes that the impenetrability of BEs constitutes evidence that BEs operate independently of cognition. The fact that they respond to low level inputs or processes to which other systems have limited access certainly suggests that emotional states can respond to information that is not integrated with cognition. In addition, there is evidence

Isaac Wiegman
10/19/2016

that people cannot fully control facial expressions of BEs (Ekman, 1972; Friesen, 1973), suggesting that BEs are cognitively impenetrable. Overall, BEs appear to lack informational integration.

Nevertheless, the realm of the cognitive picks out not only informational states, but also includes a broader range of internal states that function as causal intermediates between stimulus and response, perception and action (Rey, 1997). Cognitive states so understood include not only informational states (such as beliefs) but also motivational states (such as desires). Moreover, questions about cognitive integration may be asked about either informational or motivational states. If so, the possibility arises that the two forms of cognitive integration are independent of one another. If so, any inference from the one to the other is invalid.

This becomes clear when we consider hunger. Hunger may very well be akin to desire (a paradigmatic case of a cognitively integrated state) in the sense that it can interact with other cognitive systems to produce flexible or novel behaviors, as when rodents take novel “short cuts” to get to a food box in a maze (Olton, 1979; Tolman, 1948). Short cut behaviors suggest that hunger is a motivational state that can incline rodents to the pursuit of an end (e.g. food consumption) by selecting from a range of different means, perhaps by interacting with informational states that relate means to ends (e.g. means-ends beliefs). Even so, hunger may be cognitively impenetrable in that it may be triggered by low level stimuli and processes (e.g. low-level detection of changes in blood sugar). Moreover, when one feels hungry, one cannot interfere with the feeling of hunger by thinking about it (e.g. by noticing

Isaac Wiegman
10/19/2016

that the amount of energy one's body has stored in fat deposits is more than enough to sustain oneself). One can even imagine that it is informationally encapsulated: it might store information (e.g. about which foods are more calorically dense) that other systems cannot directly access.

These conceptual possibilities suggest that questions concerning the integration of informational states are conceptually independent of questions concerning the integration of motivational states. Hunger may be informationally encapsulated while retaining a degree of integration as a motivational state. Wholesale encapsulation, therefore, does not follow from informational encapsulation. If this is correct, then inferences like the one Izard draws above are invalid: having non-cognitive inputs is not a reason to think that emotions operate independently of cognition. They might very well operate in concert with cognition on the output side or as motivational states. Before I raise that possibility, consider another reason to rule it out at the outset: that BEs are not integrated with propositional attitudes, including beliefs *and* desires.

Assumption 2: Integration with Beliefs and Desires is the Criterion for Cognitive

Integration

Contrary to the previous assumption, this one respects the distinction between motivational and informational integration. Nevertheless, I argue that it sets the bar for cognitive integration too high.

Isaac Wiegman
10/19/2016

To see this, consider Griffiths' (Griffiths, 1997, 2004) views on the distinction between basic and higher cognitive emotions. First, he draws on some of the same evidence as Izard to conclude that BEs are opaque and informationally encapsulated. Since they have these and other marks of modularity, Griffiths thinks BEs have "limited involvement" with higher cognitive processes, which are "...the processes in which people use the information of the sort they verbally assent to (traditional beliefs) and the goals they can be brought to recognize (traditional desires) to guide relatively long-term action and to solve theoretical problems." (Griffiths, 1997, p. 92) Here, Griffiths may be making the same faulty assumption as Izard (that informational encapsulation implies cognitive encapsulation more broadly). However, let us grant that he may have additional reasons to think that emotions are not integrated on the output side or qua motivational states.

From this, Griffiths draws a broader conclusion: that BEs are not "flexible [or] integrated with long-term, planned action" and are instead "restricted to short-term, stereotyped responses" (Griffiths, 1997, p. 241). The apparent assumption is that if BEs are not integrated with beliefs, desires and long-term planning, then the only alternative is that they are similar to fixed action patterns, being inflexible and stereotyped. Griffiths makes no explicit argument for this assumption, perhaps at the time it was widespread enough to make further argument otiose.

Nevertheless, it has become a tendentious assumption for several reasons. First, the phenomena of intelligent action are much broader than deliberate, "long-term, planned action" mediated by beliefs and desires. For instance, Ginet (1990) argues that many clear

Isaac Wiegman
10/19/2016

cases of actions (as distinct from mere behaviors, such as reflexes or fixed action patterns) are not plausibly mediated by conscious beliefs, desires or intentions: involuntarily crossing one's legs, kicking a door in anger, impulsively pulling a loose thread from one's clothes, and slamming on the brakes to avoid hitting a dog. These actions are not mere behaviors or reflexes. That is, they appear to be purposive and guided by the agent, but it is difficult to find belief-desire style explanations that render them intelligible.¹ Why not think that BEs can influence actions more akin to this variety than to "long-term, planned actions"? Griffiths never raises this question, neither does he give reason to rule out the possibility that BEs cause actions intermediate between long-term planned action and stereotyped behavioral responses.

Second, if we ask what might explain the other varieties of action that Ginet picks out, it may be that such actions are guided by other representational states, aside from conscious or verbally reportable beliefs, desires and intentions. For instance, in the last twenty years, cognitive scientists have begun to emphasize the role of unconscious or non-conceptual representational states in generating flexible and intelligent behavior (Bermúdez, 2003). Informational states aside from beliefs include perceptual representations, map-like spatial representations and representations of affordances. Motivational states aside from desires include drives, incentives and feedback mechanisms.

¹ See also Hursthouse (1991).

Isaac Wiegman
10/19/2016

The flexibility and intelligence of these representational states becomes clear when we consider animal behavior. Nonhuman animals display forms of intelligent or purposive or instrumental behavior (see e.g. Balleine & Dickinson, 1998), even while lacking linguistically mediated propositional attitudes. This suggests that instrumental behaviors in non-human animals are underwritten by a different form of cognitive integration. Consider what Susan Hurley calls *holistic flexibility*:

The holistic flexibility of intentional agency contributes a degree of generality to the agent's skills: a given means can be transferred to a novel end, or a novel means adopted toward a given end. The end or goal functions as an intervening variable that organizes varying inputs and outputs and allows a degree of transfer across contexts. (Hurley, 2003, pp. 237–38)

Where this sort of flexibility is found, it suggests that behavior is best explained with reference to informational states which represent the means available to an organism (e.g. affordances) and motivational states that represent its ends (e.g. drive states), which can interact interchangeably in order to bring about the same end by various means or to deploy a single means to bring about various ends.

Nevertheless, these informational and motivational states may sometimes lack inferential integration with beliefs and desires. Even in humans, phenomena like “blind-sight” suggest that perceptual representations can flexibly guide behavior without being integrated with verbally reportable states. That is, even though these perceptual states are not verbally reportable or consciously accessible, these informational states mediate goal-

Isaac Wiegman
10/19/2016

directed behaviors (e.g. putting a plate in a slot) rather than just reflexes and fixed action patterns (see e.g. Goodale, Milner, Jakobson, & Carey, 1991). All this suggests that Griffiths' requirements on cognitive integration are too stringent. Verbal reportability and conscious accessibility of a representational state is not necessary for such a state to influence flexible behaviors. To my knowledge there is no evidence that BEs fail to meet less stringent requirements on cognitive integration such as holistic integration.

Once the full range of representational states is expanded in this way (beyond beliefs and desires), it becomes possible that BEs have some degree of motivational integration with other representational states aside from conscious beliefs and desires to produce behaviors that are more flexible and purposive than stereotyped behaviors. Griffiths provides no reason to rule out this possibility.

4. Evidence of Integration in a Basic Emotion

In fact, there is some reason to rule it in. Consider the instinctive patterns of territorial behavior of rodents. These behaviors have been investigated in great detail using a resident-intruder experimental paradigm (for an overview, see D. C. Blanchard & Blanchard, 1984, 2003) add it Adams RRR) in which resident (who have occupied a cage or colony for a few weeks) will attack unfamiliar male intruders introduced into their cage. The attacks of the resident and the defensive maneuvers of the intruder comprise sets of stereotyped behaviors. Each attack behavior of the resident is paired with a matching defensive maneuver of the intruder. The resident adopts a set of stereotyped postures and attacks aimed at biting the

Isaac Wiegman
10/19/2016

dorsal surfaces of the intruder. On the other hand, the intruder adopts a distinctive set of stereotyped behaviors aimed at avoiding or blocking the resident's attempts to bite its back.

While these behaviors are certainly stereotyped, they are not brittle or reflexive. For instance, attacks of residents vary depending on the defensive strategy adopted by the intruder, and they seem to be governed by a motive to approach and attack that persists the entire time that the intruder is present. By contrast, the intruder rat's whole suite of behaviors seems to be governed by a persistent motive to escape and avoid.

Isaac Wiegman
10/19/2016

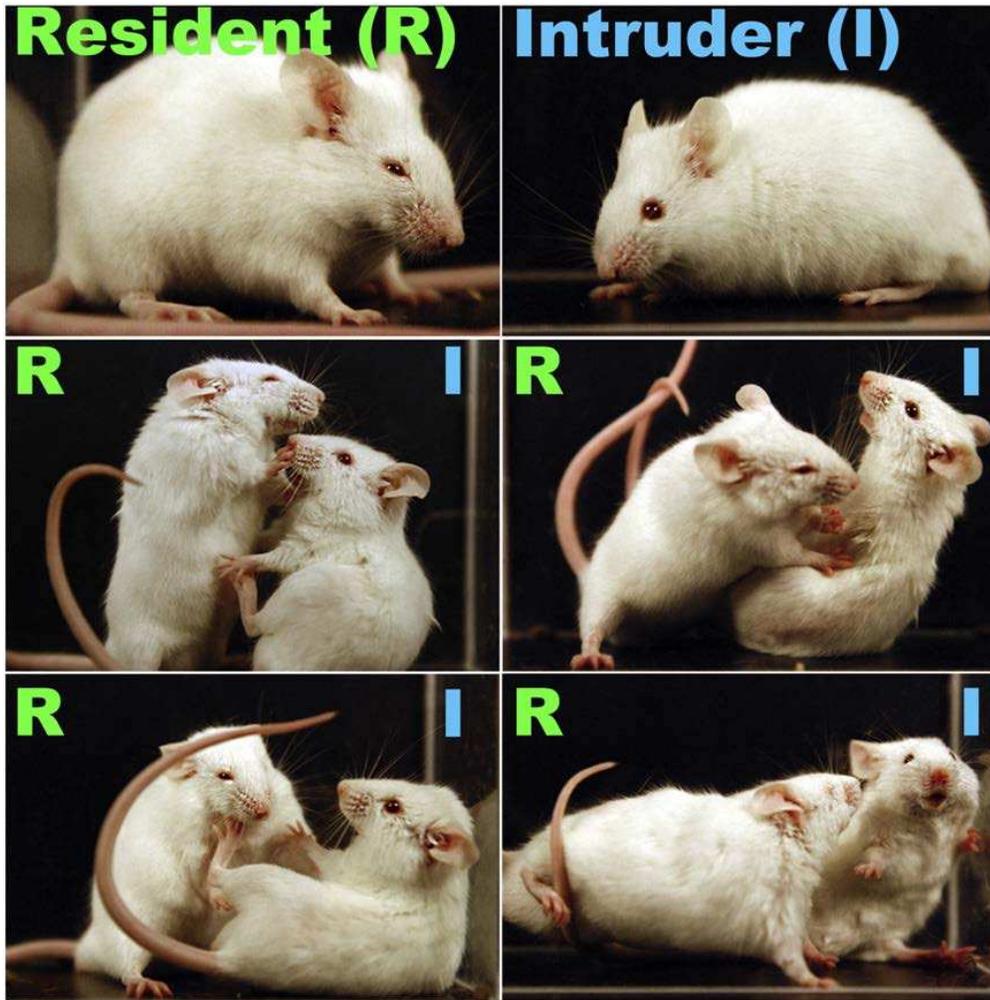


Figure 1 Confrontation and avoidance behaviors (e.g. facial expressions, postures and maneuvers) of resident and intruder mice (respectively). From Defensor and Corley (2012), p. 683 permission pending © Elsevier. Originally published in *Physiology and Behavior*.

Isaac Wiegman
10/19/2016

What scientists have discovered about these behaviors (the flexibility of these behaviors and their coherent aims) indicates that they are produced by two underlying motivational systems, what I call the confrontation and avoidance systems (D. C. Blanchard & Blanchard, 1984, 2003; D. C. Blanchard, Litvin, Pentkowski, & Blanchard, 2009). The confrontation system is tuned to bring about a specific end state, repeated back-biting. Moreover, this motive does not depend on learning: rats which have been socially isolated from birth will still attempt to bite the back of an intruder (Eibl-Eibesfeldt, 1961). So far, the focus has been on cases in which a given rodent is purely motivated by confrontation or avoidance, but aggressive encounters in the wild usually involve a mix of offensive and defensive postures. This suggests that these motivational systems can be activated simultaneously or in close succession to produce mixed patterns of behavior.

Regardless, these systems have many of the characteristics of affect programs in humans. They are posited to explain a coordinated suite of behaviors and physiological changes that may include facial expressions, cardiovascular changes, and endocrine responses (Defensor, Corley, Blanchard, & Blanchard, 2012; Fokkema, Koolhaas, & van der Gugten, 1995). Moreover, these systems are tailored to solve basic life problems. Specifically, the confrontation system solves the problem of defending territories from other males for breeding purposes (and without fatally injuring kin in the process), whereas the avoidance system solves the problem of avoiding occupied territories and failing that, defending against the attacks of residents. For these reasons, we have all the same reasons to

Isaac Wiegman
10/19/2016

postulate BEs in rodent that we have in humans. Let us suppose then that the confrontation and avoidance systems are BEs in rodents.

Interesting for my purposes, under certain conditions, the presence of the unfamiliar male can produce highly flexible and novel behaviors. In the bound-intruder task, an intruder is tied down on a Plexiglas plate with only its ventral surfaces (belly-side) exposed and placed in the cage of a resident, so that the resident cannot easily bite the back of the intruder. As a result, the resident will sometimes bite at the bands that tie down the intruder or dig under the intruder so that the resident can bite the intruder's back (R. J. Blanchard, Blanchard, Takahashi, & Kelley, 1977). In contrast, none of these behaviors are adopted when the intruder is tied down with his back exposed.

These instrumental behaviors are clearly not stereotyped forms of attack, rather they are forms of flexible behavior adjustment to achieve the aim of biting the intruder's back: they exhibit holistic integration. In this case, the same end can be achieved by several, novel means. Attempts to bite the intruder's bonds or to dig underneath the intruder are novel means toward the end of biting the back of the intruder. Moreover, some of a resident's means can be deployed toward novel ends. Digging is an element of the rat's behavioral repertoire that is ordinarily used for an entirely different purpose: constructing burrow systems for shelter and nesting (Boice, 1977). This suggests that there are informational states, representations of means (e.g. motor representations of digging, biting, lateral attack, etc.), that can interact interchangeably with motivational states, representations of various ends (e.g. nesting, back-biting, eating etc.), in order to produce flexible behaviors.

Isaac Wiegman
10/19/2016

Importantly, the confrontation system seems to be involved in coordinating flexible back-biting behavior. Moreover, this is something we would predict if it is a solution to the basic life problem of defending a territory from intruders. Flexibility is required to successfully repel an intruder because it is not in the intruder's best interest to be repelled easily or to act predictably. For instance, the intruder would be sure to fare poorly if it acted in a way that accommodates the attacks of the resident. So a single fixed action pattern or even a whole suite of fixed action patterns on the part of the resident would not tend to be successful against the most likely strategy of the intruder. It is more adaptive to have a flexible motivational state that leads to repeated back biting across a wide range of strategies or postures that the intruder might adopt. Rather than leading only to inflexible, stereotyped responses, it appears that solutions to basic life problems sometimes require some degree of motivational integration.

5. Implications for Emotion Theory

If we understand BEs in this way, this changes the shape of an ongoing debate in emotion theory concerning the existence of BEs in humans. In the past, this debate has carried on under the assumption that if an emotion is biologically basic, then one should predict that the various response components of the emotion will have a high degree of coherence; that for example "all instances of anger should have a characteristic facial display, cardiovascular pattern, and voluntary action that are coordinated in time and correlated in intensity."

Isaac Wiegman
10/19/2016

(Barrett, 2006, p. 29) This high degree of coherence is not observed across many emotions (Gentsch, Grandjean, & Scherer, 2013; Reisenzein, Studtmann, & Horstmann, 2013). For instance, when anger is elicited in experimental settings, it is uncommon to observe facial expressions in conjunction with the other putative components of BE anger.

One way of defending the basicity of an emotion against this criticism is to reassess what patterns of emotional response are predicted by BE theory. As we saw in the section above the motivational component of a basic emotion can select novel, instrumental behaviors. Moreover, the motivational component can be indispensable for solving a basic life problem. I think we can add to this the possibility that other response components are not as indispensable as the motivational state. To see this, suppose that anger in humans is a solution to basic life problems of deterring conspecifics from challenges and insults. If so, it may be that the only reliable requirement of successful deterrence (at least in our lineage) is a flexible motivation to retaliate against perceived wrongs (e.g. McCullough, Kurzban, & Tabak, 2012). For instance, a reliable disposition to garner a reputation for revenge (e.g. by avenging personal offenses) appears to be a highly reliable strategy for deterrence (e.g. Daly & Wilson, 1988; Frank, 1988), perhaps more so than any facial expression or physiological responses. If revenge can be served cold, then anger may not always require anything more than a motivation to avenge. If so, then we might *expect* that the only reliably occurring component of anger is the relevant motivational state. But if this is correct, then evidence of low coherence is not evidence against the existence of BE anger. While this is a just-so story that may or may not end up being true, it shows that the expected level of coherence in a BE

Isaac Wiegman
10/19/2016

depends on which basic life problem shaped that emotion. In some cases, we might expect the motivational state to be the only component that does not significantly vary across the situations in which these problems arise. In that case, contextually variable responses will be the norm rather than the exception.

6. Conclusion: What Basic Emotions Really Are

So what are basic emotions? Like other theoretical terms, part of the theoretical function of basic emotions is to place selective stress on competing theories (e.g. Koonin, 1985). In this case, BEs and competing conceptions of emotion allow us to discriminate between evolutionary theories of emotion in competition with radical social constructivist theories (e.g. Barrett, 2014; Lindquist, Siegel, Quigley, & Barrett, 2013).

BEs help distinguish these theories by specifying an architecture for emotion production predicted by evolutionary considerations. The distinguishing factor is whether emotion production is categorical or dimensional (see figure 2). If each BE is a solution to a different basic life problem, then when a BE is elicited, we should see emotional responses that are relevant to that basic life problem and distinct from the responses manifested by other BEs. Emotion production is categorical in the sense that the behavioral responses are controlled by a single emotional state (as distinct from other emotional states that might control a distinct pattern of response). By contrast, if all emotions are socially constructed as

Isaac Wiegman
10/19/2016

some theorists claim, we might expect to see emotional behaviors controlled directly by multiple dimensions of appraisal (as in the bottom half of figure 2).

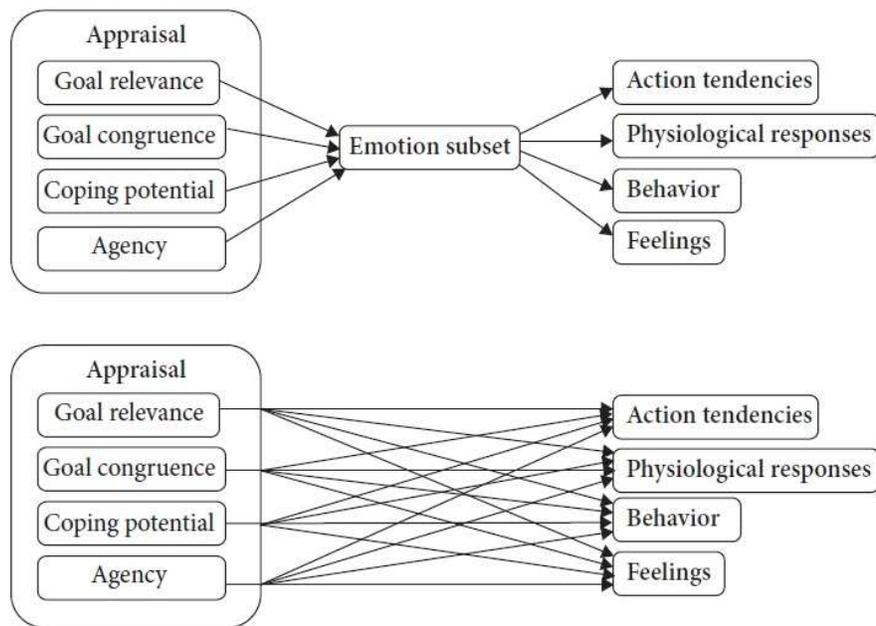


Figure 2 Competing architectures for emotion production. Top diagram is a categorical architecture, whereas the bottom is dimensional. From Moors (2012), p. 266 permission pending © John Benjamins Publishing Company. Originally published in Zachar and Ellis (2012).

Isaac Wiegman
10/19/2016

Until the present, contextual variability of emotional responses has played a decisive role in distinguishing between these two architectures for emotion production. If flexible motivational states are not included among the components of BEs, then discrete emotion production predicts insensitivity to context subsequent to elicitation (though emotion regulation processes can perhaps inhibit or augment emotional responses according to context). However, once flexible motivational states are possible, categorical emotion production is compatible with a greater amount of contextual variability.

Admittedly, this added complexity makes it more difficult to test whether humans have BEs. Nevertheless, it is not impossible. For instance, in the case of anger, researchers have developed a neurological measure of approach motivation (for a review, see Carver & Harmon-jones, 2009). If this motivational state is a component of anger, we can measure whether approach motivation itself is better predicted by contextual variables subsequent to anger elicitation or rather by contextual variables prior to or during elicitation. If contextual variables prior to elicitation do not independently predict approach motivation as BE theory might lead us to expect, then we would have evidence against the existence of BE anger.

I have argued against prevailing assumptions that BEs lack cognitive integration. In the past, evidence against cognitive integration has been concerned with informational integration, and motivational integration has not been considered. Moreover, the assumed requirements for integration concern interaction with verbally reportable or consciously accessible states, and integration with other representational states is ignored. Moreover, BEs in rodents exhibit a form of motivational integration that plausibly hinges on interaction with

Isaac Wiegman

10/19/2016

a wider variety of representational states. Properly understood, BEs are more likely to refer to emotional states in humans.

Isaac Wiegman
10/19/2016

References

- Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407–419.
- Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science*, *1*(1), 28–58. <http://doi.org/10.1111/j.1745-6916.2006.00003.x>
- Barrett, L. F. (2014). The Conceptual Act Theory: A Précis. *Emotion Review*, 1–20. <http://doi.org/10.1177/1754073914534479>
- Bermúdez, J. (2003). *Thinking without words*.
- Blanchard, D. C., & Blanchard, R. J. (1984). Affect and aggression: An animal model applied to human behavior. In R. J. Blanchard & D. C. Blanchard (Eds.), *Advances in the Study of Aggression* (Vol. 1, pp. 1–62).
- Blanchard, D. C., & Blanchard, R. J. (2003). What can animal aggression research tell us about human aggression? *Hormones and Behavior*, *44*(3), 171–177. [http://doi.org/10.1016/S0018-506X\(03\)00133-8](http://doi.org/10.1016/S0018-506X(03)00133-8)
- Blanchard, D. C., Litvin, Y., Pentkowski, N. S., & Blanchard, R. J. (2009). Defense and Aggression. In G. G. Berntson & J. T. Cacioppo (Eds.), *Handbook of Neuroscience for the Behavioral Sciences* (pp. 958–974). Hoboken: Wiley.
- Blanchard, R. J., Blanchard, D. C., Takahashi, T., & Kelley, M. J. (1977). Attack and defensive behaviour in the albino rat. *Animal Behaviour*, *25*, 622–634.

Isaac Wiegman
10/19/2016

- Boice, R. (1977). Burrows of wild and albino rats: effects of domestication, outdoor raising, age, experience, and maternal state. *Journal of Comparative and Physiological Psychology*, 91(3), 649–61.
- Carver, C. S., & Harmon-jones, E. (2009). Anger Is an Approach-Related Affect : Evidence and Implications. *Psychological Bulletin*, 135(2), 183–204.
<http://doi.org/10.1037/a0013965>
- Daly, M., & Wilson, M. (1988). *Homicide*. Transaction Publishers.
- Defensor, E. B., Corley, M. J., Blanchard, R. J., & Blanchard, D. C. (2012). Facial expressions of mice in aggressive and fearful contexts. *Physiology & Behavior*, 107(5), 680–5. <http://doi.org/10.1016/j.physbeh.2012.03.024>
- Eibl-Eibesfeldt, I. (1961). The Fighting Behavior of Animals. *Scientific American*, 205, 112–122.
- Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*. University of Nebraska Press Lincoln.
<http://doi.org/10.1037/0022-3514.53.4.712>
- Ekman, P. (1977). Biological and cultural contributions to body and facial movement. In J. Blacking (Ed.), *Anthropology of the body* (pp. 34–84).
- Ekman, P. (1999). Basic Emotions. In T. Dalgleish & M. Power (Eds.), *The Handbook of Cognition and Emotion* (pp. 45–60). Sussex: John Wiley & Sons.

Isaac Wiegman
10/19/2016

- Ekman, P. (2003). *Emotion Revealed: Understanding Faces and Feelings*. Phoenix Press.
- Ekman, P., & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review*, 3(4), 364–370. <http://doi.org/10.1177/1754073911410740>
- Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616), 1208–1210.
- Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*.
- Fokkema, D. S., Koolhaas, J. M., & van der Gugten, J. (1995). Individual characteristics of behavior, blood pressure, and adrenal hormones in colony rats. *Physiology & Behavior*, 57(5), 857–62.
- Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. New York: Norton. <http://doi.org/10.2307/2072516>
- Friesen, W. (1973). *Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules*. University of California, San Francisco.
- Gentsch, K., Grandjean, D., & Scherer, K. R. (2013). Coherence explored between emotion components: Evidence from event-related potentials and facial electromyography. *Biological Psychology*. <http://doi.org/10.1016/j.biopsycho.2013.11.007>
- Ginet, C. (1990). *On action*.
- Goodale, M., Milner, A., Jakobson, L., & Carey, D. (1991). A neurological dissociation

Isaac Wiegman
10/19/2016

between perceiving objects and grasping them. *Nature*.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories* (Vol. 1997). University of Chicago Press.

Griffiths, P. E. (2004). Emotions as Natural and Normative Kinds, *71*(December), 901–911.

Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language*, *18*(3), 231–256.

Hursthouse, R. (1991). Arational actions. *The Journal of Philosophy*.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations.

Izard, C. E. (2007). Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280. <http://doi.org/10.1111/j.1745-6916.2007.00044.x>

Kroon, F. (1985). Theoretical terms and the causal view of reference. *Australasian Journal of Philosophy*, (February 2014), 37–41.

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*(4), 363–384.

Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2013). The Hundred-Year Emotion War : Are Emotions Natural Kinds or Psychological Constructions ? Comment on Lench , *139*(1), 255–263. <http://doi.org/10.1037/a0029038>

Isaac Wiegman
10/19/2016

- McCullough, M. E., Kurzban, R., & Tabak, B. a. (2012). Cognitive systems for revenge and forgiveness. *The Behavioral and Brain Sciences*, 1–15.
<http://doi.org/10.1017/S0140525X11002160>
- Moors, A. (2012). Comparison of affect program theories, appraisal theories, and psychological construction theories. *Categorical versus Dimensional Models of Affect. A Seminar on the Theories of Panksepp and Russell*, 257–278.
- Olton, D. (1979). Mazes, maps, and memory. *American Psychologist*.
- Reisenzein, R., Studtmann, M., & Horstmann, G. (2013). Coherence between Emotion and Facial Expression: Evidence from Laboratory Experiments. *Emotion Review*, 5, 16–23.
<http://doi.org/10.1177/1754073912457228>
- Rey, G. (1997). Contemporary philosophy of mind: A contentiously classical approach.
- Tolman, E. (1948). Cognitive maps in rats and men. *Psychological Review*.
- Zachar, P., & Ellis, R. (2012). *Categorical versus dimensional models of affect: a seminar on the theories of Panksepp and Russell*.

Multiple realization and the commensurability of taxonomies

Abstract

The past two decades have witnessed a revival of interest in multiple realization and multiply realized kinds. Bechtel and Mundale's (1999) illuminating discussion of the subject must no doubt be credited with having generated much of this renewed interest. Among other virtues, their paper expresses what seems to be an important insight about multiple realization: that unless we keep a consistent grain across realized and realizing kinds, claims alleging the multiple realization of psychological kinds are vulnerable to refutation. In this paper I argue that, intuitions notwithstanding, the terms in which their recommendation has been put make it impossible to follow, while also misleadingly insinuating that meeting their desideratum virtually guarantees mind-brain identity. Instead of a matching of grains, what multiple realization really requires is a principled method for adjudicating upon differences between tokens. Shapiro's (2000) work on multiple realization can be understood as an attempt to adumbrate such a method.

Multiple realization, neuroscience, autonomy of psychology, intertheoretic reduction

1. Introduction

The multiple realization (“MR”) hypothesis asserts, at its baldest, that the same psychological state may be realized in neurologically distinct substrates (Polger 2009). Hilary Putnam’s (1967) ingenious suggestion that pain is likely to be a multiply realized kind (“MR kind”) rather neatly captures the thought here—while both mammals and molluscs presumably experience pain, they’re likely to instantiate it in neurological systems of a very different sort.

MR was played against a popular philosophical theory of mind in the 1960s which attempted to identify mental states with neural states. Since MR implies a many-to-one mapping from neural states to mental states, if it is in fact true that mental states are multiply realized, it follows that no clear identity relation can hold between them. As Bechtel and Mundale (1999, 176) frame the issue, “[o]ne corollary of this rejection of the identity thesis is the contention that information

about the brain is of little or no relevance to understanding psychological processes." When the MR hypothesis first came to prominence, its critics by and large accepted it as empirically correct, and merely denied its touted antireductionist implications. In recent years the debate has struck a new note, with many philosophers calling the empirical hypothesis itself into question. Bechtel and Mundale's (1999) influential paper, followed quickly at the heels by Shapiro's (2000) penetrating analysis of functions, perhaps did most to reignite the old controversy and drag MR back into the philosophical limelight. Bechtel and Mundale express what seems to be an important insight about multiple realization: that unless we keep a consistent grain across realized and realizing kinds, claims alleging the multiple realization of psychological kinds are vulnerable to refutation. In this paper I argue that, intuitions notwithstanding, the terms in which their recommendation has been put make it impossible to follow, while also misleadingly insinuating that meeting their desideratum virtually guarantees mind-brain identity. Instead of a matching of grains, what MR really requires is a principled method for adjudicating upon differences between tokens. Shapiro's (2000) work on MR can be understood as an attempt to adumbrate such a method.

2. Bechtel and Mundale's grain requirement

Bechtel and Mundale appeal to “neurobiological and cognitive neuroscience practice” in the hope of showing how claims that psychological states are multiply realized are unjustified. Intuitively, theirs is an argument from success: cognitive neuroscience’s method assumes MR is false, and the success of that method is evidence that MR *is* false. They argue that it is “precisely on the basis of working assumptions about commonalities in brains across individuals and species that neurobiologists and cognitive neuroscientists have discovered clues to the information processing being performed” (1999, 177).

Bechtel and Mundale examine both the “neuroanatomical and neurophysiological practice of carving up the brain.” What they believe this examination reveals is, firstly, that the principle of psychological function plays an essential role in both disciplines, and secondly, that “the cartographic project itself is frequently carried out comparatively—across species” (1999, 177), the opposite of what one would expect if MR were “a serious option.” It is the very similarity (or homology) of brain structure which permits generalization across species; and similarity in the functional characterization of homologous brain regions across

species only makes sense if the claims of MR are either false or greatly exaggerated. For instance, “[e]ven with the advent of neuroimaging, permitting localization of processing areas in humans, research on brain visual areas remains fundamentally dependent on monkey research...” (1999, 195). “The clear assumption is that the neural organization in the macaque will provide a defeasible guide to the human brain” (1999, 183). Brodmann’s famous brain maps were based upon comparisons of altogether 55 species and 11 orders of mammals. If MR were true, “one would not expect results based on comparative neuroanatomical and neurophysiological studies to be particularly useful in developing functional accounts of human psychological processing” (1999, 178). They also argue that the ubiquity of brain mapping as a way of decomposing cognitive function points to the implausibility of the MR thesis. The understanding of psychological function is increasingly “being fostered by appeal to the brain and its organization” (1999, 191), again, the opposite of what one would expect “[i]f the taxonomies of brain states and psychological states were as independent of each other as the [MR] argument suggests” (1999, 190-91).

In light of such considerations, Bechtel and Mundale (1999, 178-79, 201-04) resort to grains as a way of making sense of what they perceive to be the

entrenched, almost unquestioning consensus prevailing around MR. They think that it can be traced to the practice of philosophers appealing to different grain sizes in the taxonomies of psychological and brain states, “using a coarse grain in lumping together psychological states and a fine grain in splitting brain states.”

When Putnam went about collecting his various specimens of pain, he ignored the many likely nuances between them. At the same time, he had few compunctions about declaring them different at a neurological level. His contention that pain is likely to be an MR kind can only command our respect if we can be sure that when he was comparing his specimens from a neurological point of view he was careful to apply no less lenient a standard of differentiation than he applied when comparing his specimens from a psychological point of view. Bechtel and Mundale maintain that when “a common grain size is insisted on, as it is in scientific practice, the plausibility of multiple realizability evaporates.” As their examples of neuroanatomical and neurophysiological practice attest, scientists in these fields typically match a coarse-grained conception of psychological states with an equally coarse-grained conception of brain states. Despite the habit of philosophers individuating brain states in accordance with physical and chemical criteria, a habit no doubt originating with Putnam, this is not how neuroscientists characterize them. The notion of a brain state is “a philosopher’s fiction” (1999,

177) given that the notion neuroscientists actually employ is much less fine-grained, namely “activity in the same brain part or conglomerate of parts.”

A not unrelated factor is that the MR hypothesis often gets presented in a “contextual vacuum.” The choice of grain is always determined by context, with “different contexts for constructing taxonomies” resulting in “different grain sizes for both psychology and neuroscience.” The development of evolutionary perspectives, for instance, in which the researcher necessarily adopts a coarse grain, contrasts with the much finer grain that will be appropriate when assessing differences among conspecifics:

One can adopt either a coarse or a fine grain, but as long as one uses a comparable grain on both the brain and mind side, the mapping between them will be correspondingly systematic. For example, one can adopt a relatively coarse grain, equating psychological states over different individuals or across species. If one employs the same grain, though, one will equate activity in brain areas across species, and one-to-one mapping is preserved (though perhaps further taxonomic refinement and/or delineation may be required). Conversely, one can adopt a very fine grain,

and differentiate psychological states between individuals, or even in the same individual over time. If one similarly adopts a fine grain in analyzing the brain, then one is likely to map the psychological differences onto brain differences, and brain differences onto psychological differences. (1999, 202)

At least among some philosophers Bechtel and Mundale's message has evidently been well received (Couch 2004; Polger 2009; Godfrey-Smith, personal communication; see also tacit approval in Aizawa and Gillett 2009, 573). Polger (2009) explains the motivation for the grain requirement in an illuminating way. Neuroplasticity has in recent times been thought to provide compelling evidence for the MR of mental states. He concludes that "contrary to philosophical consensus, the identity theory does not blatantly fly in the face of what is known about the correlations between psychological and neural processing" (2009, 470). The grains argument figures prominently in his reasoning. As he points out, it might be tempting to regard a phenomenon like cortical map plasticity—where different brain regions subserve the same function at different times in an individual's history, say, after brain injury or trauma—as an existence proof of MR. But not if the point about grains is taken to heart. It all comes down to what we mean by "*different* brain regions" subserving "*the same* function." Consider that

recovered functions are frequently suboptimal. Genuine MR would indeed require the *same* psychological state to be underwritten by different neurological states; but suboptimality is evidence of difference underlying difference, not difference underlying sameness, as MR requires:

It's true that this kind of representational plasticity involves the "same" function being mediated by "different" cortical areas. But here one faces the challenge leveled by Bechtel and Mundale's charge that defenses of [MR] employ a mismatch in the granularity of psychological and neuroscientific kinds. If we individuate psychological processes quite coarsely—by gross function, say—then we can say that functions or psychological states are of the same kind through plastic change over time. And if we individuate neuroscientific kinds quite finely—by precise cortical location, or particular neurons—then we can say that cortical map plasticity involves different neuronal kinds. But this is clearly a mug's game. What we want to know is not whether there is some way or other of counting mental states and brain states that can be used to distinguish them—no doubt there are many. The question is whether the sciences of psychology and neuroscience give us any way of *registering the two taxonomic systems*. (2009, 467, my emphasis)

3. Problems with the grain requirement: imprecise, impracticable, and misleading

But now the question is this: what, precisely, can it mean to use a “comparable” grain, or to keep a grain size “constant,” across both psychological and neurophysiological taxonomies? Polger’s motivation makes a lot of sense, to be sure, but talk of “registering” taxonomies (as of *aligning* classificatory regimes, or rendering distinct scientific descriptions *commensurable*, or however else one might care to put it) doesn’t shed any light on how the desideratum for consistent grains can actually be met. Since it is intended to serve in part as a methodological prescription, it’s important to know what to make of this requirement—metaphors won’t help us here. How, in *concrete* terms, is an investigator meant to satisfy such a condition as *this* on their research?

Perhaps it means this. Suppose you have two tokens of fruit. The science of botany (say) could deliver descriptions under which the two are classified the same (e.g. from the point of view of *species*), but also descriptions under which they come out as different (e.g. from the point of view of *varieties*). The first

description could be said to apply a coarser grain than the second. Now imagine economics coming into the picture. The science of economics can likewise deliver descriptions under which both tokens are classified the same (e.g. both are forms of tradable fresh produce) or different (e.g. one, being typically the crunchier and sweeter variety, has a lower elasticity of demand than the other). Once again, the first description could be said to apply a coarser grain than the second. Perhaps, then, we could take it that botany and economics deliver descriptions at the same grain of analysis when their judgments of sameness or difference cohere in a given case. In the example, botanical descriptions via species classification would be furnished at the same grain as economic descriptions via commodity classification, so that species descriptions in botany are “at the same grain” as commodity descriptions in economics. By the same logic, *variety* descriptions in botany would be comparable to *elasticity* descriptions in economics. Fine. But if that is all that “maintain a comparable grain” amounts to, it really does beg the question, for this is simply type-type identity by fiat. *Of course* such a recommendation will ensure that the mapping between psychology and neuroscience will be “systematic” (to use Bechtel and Mundale’s term), because on this account yielding concordant judgments of similarity or difference across taxonomies is what it *means* to apply the same grain. So we haven’t solved the problem: *this* version of the grain

requirement makes type-type identity a fait accompli, effectively obliterating all MR kinds from the natural order.

It's just as well that I don't think this is what Bechtel and Mundale had in mind when they made their move to grains; supposing otherwise would serve only to trivialize an important aspect of their analysis. Still the construal is by no means far-fetched: "[o]ne can adopt either a coarse or a fine grain," they tell us, "but as long as one uses a comparable grain on both the brain and mind side, the mapping between them will be correspondingly systematic" (note that—it *will* be!). This sounds like someone with the utmost confidence in the grain requirement, which is of course what one *would* have if one thought grains could be legitimately matched in just this way. My guess is that, while they do have something important to tell us about MR, a beguiling metaphor has led them to suppose that MR is easier to refute than it actually is. (I'll support this contention with a few examples in a moment.)

Of course matters aren't much helped by the reasonable suspicion that MR is the result of pairing *inconsistent* grains. For what is neuroscience if not a fine-grained description of psychology, and psychology if not a coarse-grained

description of neuroscience? It is surely plausible that the neural and psychological sciences line up in something like this way, given that talk about the mind is really talk about the brain from a somewhat more abstract point of view.

What Bechtel and Mundale are ultimately trying to convey through their discussion of grains is the thought that claims of MR cannot be advanced willy-nilly—that there is an objective and standard way to go about verifying the existence of MR kinds and arbitrating disputes involving them. For the reasons just canvassed, however, it strikes me that talk of grains doesn't serve their purposes at all well. In fact they would have been nearer the mark had they said that what MR requires is some sort of principled *mismatching* of grains.

So far I've tried to indicate in what respects Bechtel and Mundale's grain requirement is imprecise and impracticable. Before I can show that the grains strategy is also misleading, and indeed often gets things wrong, I need to set it against an account which demonstrably gets things right.¹ Shapiro (2000) expresses with enviable lucidity what I think is the crucial insight towards which Bechtel

¹ It is an account which even its detractors concede gets at least the essential point of interest to us here right, e.g. Gillett (2003).

and Mundale were uneasily groping. Interestingly, some philosophers—e.g. Polger (2009)—write as if the grain requirement and Shapiro’s own formula for MR were effectively interchangeable. This is a mistake: the two approaches deliver different judgments in nontrivial cases (as I’ll illustrate in a moment).

As Shapiro reminds us:

Before it is possible to evaluate the force of [the MR thesis] in arguments against reductionism, we must be in a position to say with assurance what the satisfaction conditions for [the MR thesis] actually are. (2000, 636)

For him, “[t]he general lesson is this. Showing that a kind is multiply realizable, or that two realizations of a kind are in fact distinct, requires some work” (2000, 645).

Furthermore, “[t]o establish [the MR thesis], one must show that the differences among purported realizations are causally relevant differences” (2000, 646).

Shapiro’s concerns revolve around what motivates ascriptions of difference, and therefore sameness. The issue is important because the classic intuition pump that asks us to conceive a mind in which every neuron has been replaced by a silicon chip depends on our ascription of an interesting difference between neurons and

silicon chips, apparently even where silicon chips can be made that contribute to psychological capacity by one and the same process of electrical transmission. His answer too, like Bechtel and Mundale's, depends ultimately on context—in particular, the context set by the very inquiry into MR itself.

Shapiro (2000, 643-44) argues that “the things for which [the MR thesis] has a chance of being true” are all “defined by reference to their purpose or capacity or contribution to some end.” This is the reason why carburetors, mousetraps, computers and minds are standard fare in the literature of MR. They are defined “in virtue of what they do,” unlike, say, water, which is typically defined by what it is, i.e. its constitution or molecular structure, and accordingly *not* an MR kind. Genuine MR requires that there be “*different ways* to bring about the function that defines the kind.” Truly distinct (indeed *multiple*) realizations are those that “differ in causally relevant properties—in properties that make a difference to how [the realizations] contribute to the capacity under investigation.” Two corkscrews differing only in color are not distinct realizations of a corkscrew, because color “makes no difference to their performance as a corkscrew.” Similarly, the difference between steel and aluminium is not enough to make two corkscrews that are alike in all other respects two different realizations of a corkscrew “because, relative to

the properties that make them suitable for removing corks, they are identical." In this instance, differences of composition can be "screened off." Naturally there may be cases where differences of composition *will* be causally relevant (and it turns out that this will be important to the broader point I make below about where the grains strategy goes wrong). Perhaps rigidity is the allegedly MR kind in question. In that event, compositional differences will necessarily speak to how aluminium and steel achieve this disposition. The crucial thing to note here is that MR *is* the context, and MR makes *function* the relevant consideration, i.e. the specific point of view from which we will compare a set of tokens in the first instance (not phenomenology, not behavioral ecology, or anything else for that matter). Explanatory considerations may of course fine-tune the *sort* of function that captures our attention (cork-removal, rigidity, vision, camera vision, etc.). But function here is our key preoccupation, and having settled on a specific function which a set of tokens can be said to perform, the all-important question on Shapiro's analysis is *how* the two tokens bring that function about. Each case must be judged on its own merits. Thus unlike the two corkscrews identical in all respects save color, which do not count as distinct realizations, waiter's corkscrews and winged corkscrews are enabled to perform the same task in virtue of *different* causally relevant properties, and therefore *do* count as genuinely distinct

realizations of a corkscrew, one based on the principle of simple leverage, the other relying on a rack and pinions (Fig. 1).

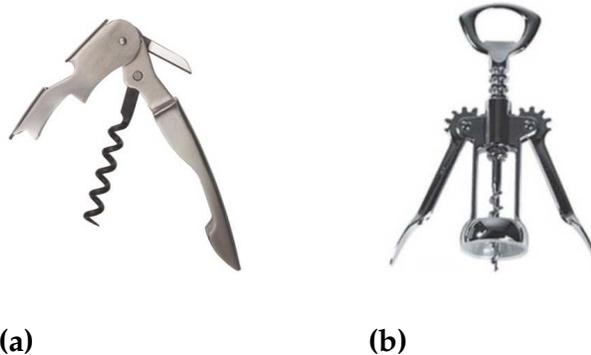


Figure 1. A waiter's corkscrew (a) and a winged corkscrew (b). Each contributes to the capacity of cork-removal in different ways.

Notice that to the extent Shapiro's causal relevance criterion envisages certain realizing properties being "screened off" from consideration in the course of inquiry, there is a sense in which the taxonomies of realized and realizing kinds may be said to be "commensurable" or "registrable" (no doubt explaining why some philosophers have simply confused commensurability with causal relevance). Thus when comparing the cork-removing properties of two waiter's corkscrews, compositional differences will not feature in the realizing taxonomy (if we accept Shapiro's characterization of the problem). So we have *cork-removal*,

which features in what we may regard as a coarse-grained taxonomy, realized by two objects described by a “science” of cork-removal in which microstructural variations do not matter, hence which might also be regarded as a coarse-grained taxonomy. If on the other hand we were comparing the same corkscrews for rigidity, where one was made of steel and the other of aluminium, compositional differences *would* feature in the realizing taxonomy. Here we would have *rigidity*, which features in what we could well regard as a more fine-grained taxonomy than that encompassing cork-removal, realized by two objects described by a science in which microstructural variations really *do* matter (namely metallurgy), and which might also be regarded as a fine-grained taxonomy, at least more fine-grained than the fictitious science of cork-removal. But my point is this: commensurability nowhere appears as an independent criterion of validity in Shapiro’s account of MR, for it is an artifact of the causal relevance criterion, not a self-standing principle. Taxonomic commensurability is in fact an *implicit* requirement of the causal relevance criterion in the sense that it’s taken care of once the proper question is posed. As an explicit constraint it is a will-o’-the-wisp.

Armed with this analysis, let’s examine how Bechtel and Mundale attempt to refute the status of hunger as an MR kind. Putnam (1967) had compared hunger

across species as diverse as humans and octopuses to illustrate the likelihood that some psychological predicates are multiply realizable. On the basis of their grains critique, however, Bechtel and Mundale suggest that hunger will not do the work Putnam had cut out for it; for “at anything less than a very abstract level,” hunger is different in octopuses and humans (1999, 202). The thought is that a finer individuation of hunger refutes the existence of a *single* psychological kind, hunger, which can be said to cross-classify humans and octopuses. Thus they essay to challenge the cognitive uniformity which MR requires at the level of psychology.

Perhaps we might first note that when identifying a *single* psychological state to establish the necessary conditions for MR, nothing Bechtel and Mundale say actually *precludes* the choice to go abstract. If context is what fixes the choice of grain (as they are surely right to point out), who’s to say that context couldn’t fix the sort of grain that makes hunger relevant in an abstract sense? It may be tempting to think that a more detailed description of something is somehow more *real*. But there is of course nothing intrinsically more or less real about a chosen schema relative to others that might have been chosen. There is no reason to suspect, for instance, that a determinate has any more reality than a determinable.

And yet there is a deeper problem with Bechtel and Mundale's deployment of the grains strategy here. To repeat their complaint: "at anything less than a very abstract level," hunger is different in octopuses and humans. But now why should *this* be relevant? Who would deny it? They themselves seem to be oblivious to the context which the very inquiry into MR makes paramount. They are not right to allege, as they do, that "the assertion that what we broadly call 'hunger' is the same psychological state when instanced in humans and octopi has apparently been widely and easily accepted without specifying the context for judging sameness" (1999, 203). The reason why hunger, pain, vision and so on were all taken for granted—assumed to be uniform at the cognitive level—is because MR made *function* the point of view from which tokens were to be compared. As Shapiro reminds us, "the things for which [the MR thesis] has a chance of being true" are all "defined by reference to their purpose or capacity or contribution to some end." It was understood that, say in the case of pain, regardless of phenomenal, ecological or behavioral differences between human and octopus pain (I doubt any of which were lost on Putnam), all instances of pain in these creatures had something like *detection and avoidance* in common. This might be to cast pain at "a very abstract level," but this just happens to be the context which

the inquiry into MR itself sets. A similarly abstract feature is what unites all instances of hunger: let's call it *nutrition-induction*. It is not that decades of philosophers had simply forgotten to specify the point of view from which these psychological predicates were being considered: it is rather that they simply didn't need to, since all of them had read enough of Putnam and the early functionalists to know what they were about. Phenomenal and other differences that one might care to enumerate between these predicates come a dime a dozen. But the whole point of functionalism was to abjure the inquiry into essences and focus instead on the causal role of a mental state within the life of an organism. Yes, this is to compare tokens from an "abstract level," but that's what made functionalism intriguing to begin with. And if Shapiro's analysis is any guide, it is really the *next* step in the endeavor to verify the existence of an MR kind that is the crucial one. Genuine MR requires that there be "*different ways to bring about the function that defines the kind.*" So the follow-up question concerns *how* the relevant organisms achieve their detection and avoidance function, or nutrition-induction function, or whatever the case may be. It is in fact only by asking this next question that we can appreciate just how badly the grains strategy fares. The attempt to individuate hunger more finely does *not* refute the multiple realizability of hunger as between humans and octopuses. For, relative to the shared function of nutrition-induction,

it is extremely likely that humans and octopuses realize this capacity in different ways. The attempt to individuate pain more finely would likewise *not* refute the multiple realizability of pain as between humans and octopuses. For, relative to the shared function of detection and avoidance, it is extremely likely that humans and octopuses realize this capacity in different ways. So we see that the grains strategy, to the extent that it involves fine-graining psychological states in order to undermine the cognitive uniformity required by MR, sets itself a very easy job indeed, and mischaracterizes the nature of MR by its neglect of function. Moreover Shapiro's causal relevance criterion—which honors the core concerns motivating Bechtel and Mundale's resort to grains—does *not* demonstrate that hunger (or pain) is type-reducible.

A good illustration of the grains strategy in action is provided by Couch's (2004) attempt to refute the claim that the human eye and the octopus eye are distinct realizations of the kind *eye*. Conceding differences at a neurobiological level, the strategy again involves challenging the alleged uniformity at the cognitive level. As he explains, “[e]stablishing [MR] requires showing that...the physical state types in question are distinct [and] that the relevant functional properties are type identical. Claims about [MR] can be challenged at either step”

(2004, 202). Reminding us that psychological states “are often only superficially similar,” and that “at a detailed level the neural differences make for functional differences” (2004, 203), he states:

Psychologists sometimes talk about humans and species like octopi sharing the same psychological states. However, they also recognize that there are important differences involved depending on how finely one identifies the relevant features...Establishing multiple realization requires showing that the same psychological state has diverse realizations. But we can always disagree with the functional taxonomy, and claim there are psychological differences at another level of description. (2004, 203)

Thus he relates that while the two types of eyes have similar structure in certain respects, both consisting of a spherical shell, lens and retina, they use different kinds of visual pigments in their photoreceptors, as well as having different numbers of them, the octopus having one in contrast to the human eye which has four. They also have different retinas. The human retina, with rods and cones, focuses light by bending the lens and so changing its shape. The octopus eye, with rhabdomeres instead of rods and cones, focuses light by moving the lens

backwards and forwards within the shell. All these factors show up as differences in output, not just structure. The octopus, having only a single pigment, is colorblind, while its receptor's unique structure allows it to perceive the plane of polarized light. Retinal differences likewise make for functional differences, with very little information processing occurring on the octopus's retina, unlike the case of the human retina. This produces differences in stimuli and reaction times. So the two eyes might be similar, but when described with a suitably fine grain, he contends, they come out type distinct. In the result they are both physically *and* cognitively diverse, and so not genuine examples of MR.

Notice again that, contrary to what is claimed, it has not been demonstrated that type-type identity prevails here after at all (on the understanding that the kind camera eye_{human} reduces to *its* distinct neural type, and the kind camera eye_{mollusc} in turn reduces to *its* distinct neural type). If anything what this foray into mollusc visual physiology succeeds in showing is that, relative to the kind camera eye, human camera eyes and octopus camera eyes count as distinct realizations(!), for, assuming Shapiro's causal relevance criterion applies, human camera eyes achieve the function of *camera vision* differently to the way octopus camera eyes

achieve this function. Were we to attend to the original inquiry, which concerned whether human eyes and octopus eyes count as distinct realizations of the kind eye, Shapiro's own response, for what it's worth, is clear (2000, 645-46): here we do seem to confront a genuine case of type-type identity, as Putnam himself assumed, because, relative to the function of *vision* (not *camera vision*), both humans and molluscs achieve the function the same way (namely, by camera vision!).

Differences that would be relevant at the neural level between humans and molluscs when asking how camera vision is achieved can be conveniently screened off when the question is how vision, as distinct from camera vision, is achieved.

Again if pain or hunger were the kind in question, it seems more likely than not that we *would* confront a case of MR (unlike with vision), as we conjectured earlier.

Explanatory context dictates the function of interest, and the function is one that we have to assume is common to the tokens in question in order to get the inquiry into MR off the ground. Indeed if Shapiro's analysis is correct, with MR we're always asking how some common function is achieved by different tokens that *do that thing*. Where there is no common function the question of MR cannot so much as arise. The fact that the question *does* arise in all the cases we've considered is a powerful indication that we're dealing with functions which all the relevant tokens actually share. The grains strategy confuses matters by suggesting that in many

cases involving putative MR kinds, psychological states can be individuated using a finer grain of description. But if what I have been saying is right, this is not the proper way to refute a putative case of MR.

That mine is the correct assessment of the situation is not only attested to by Shapiro's analysis of MR, but also by the fact that it avoids the very mug's game Polger sought to eschew by embracing the grains strategy in the first place. If for any putative MR kind I am free to cavil with the choice of your size of grain ("oh, that's far too coarse for psychology," or "now that's really not coarse enough for neuroscience"), how is the resulting game any less of a mug's game than the one we were trapped in at the start? I myself have played a few of these games with philosophers. No one wins. Couch's remarks are telling: "we can always disagree with the functional taxonomy, and claim there are psychological differences at another level of description." So the game goes on.

4. Conclusion

In sum, I think there's a genuine problem with the grain requirement. The central difficulty is that in the terms in which it's been put it is largely unworkable, and at

best no more than a loose metaphor. For a recommendation intended to serve at least in part as a methodological reform, this is clearly unsatisfactory. I don't deny that Bechtel and Mundale were onto something. But whatever value their insight into MR might have been obscured by their unfortunate formulation of the issue. Moreover, as I have tried to show, the formulation is unfortunate not *just* because it happens to be unworkable. More worryingly, the argument from grains distorts the truth about MR by encouraging the view that mind-brain identity comes for free once we invoke the "same grain" of description across both realized and realizing kinds. But when the insight to which this locution seems to point is expressed in terms that are intelligible and empirically tractable (namely, Shapiro's causal relevance criterion), mind-brain identity seems anything but a *fait accompli*. Grains talk makes it tempting to think MR is easier to refute than it in fact is. It is certainly true, as Bechtel and Mundale acknowledge, that context fixes the choice of grain (where by "grain" we mean the respect under which we seek to compare a set of tokens); but we are not *ipso facto* obliged to employ a consistent grain across realized and realizing kinds (since this is just about meaningless as far as a researcher into these matters would be concerned and raises a host of difficulties beside). Rather than matching grains, what MR really behooves us to do is to apply a principled method for adjudicating upon differences between tokens of a

functional kind. Shapiro's work on MR shows us how to approach this important task.

References

Aizawa, Kenneth, and Carl Gillett. 2009. "Levels, Individual Variation, and Massive Multiple Realization in Neurobiology." In *The Oxford Handbook of Philosophy and Neuroscience*, ed. John Bickle, 539-81. New York: Oxford University Press.

Bechtel, William, and Jennifer Mundale. 1999. "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science* 66(2): 175-207.

Couch, Mark B. 2004. "A Defense of Bechtel and Mundale." *Philosophy of Science* 71(2): 198-204.

Gillett, Carl. 2003. "The metaphysics of realization, multiple realizability, and the special sciences." *Journal of Philosophy* 100(11): 591-603.

Polger, Thomas W. 2009. Evaluating the evidence for multiple realization. *Synthese* 167(3): 457-472.

Putnam, Hilary. 1967. Psychological predicates. In: *Art, mind, and religion*, eds. W. Capitan & D. Merrill, 37-48. Pittsburgh: University of Pittsburgh Press.

Shapiro, Lawrence A. 2000. "Multiple Realizations." *Journal of Philosophy* 97(12): 635-54.

Interventionist Causation in Thermodynamics

Karen R. Zwier

March 2016 (Preprint)

Abstract

The interventionist account of causation has been largely dismissed as a serious candidate for application in physics. This dismissal is related to the problematic assumption that physical causation is entirely a matter of dynamical evolution. In this paper, I offer a fresh look at the interventionist account of causation and its applicability to thermodynamics. I argue that the interventionist account of causation is the account of causation which most appropriately characterizes the theoretical structure and phenomenal behavior of thermodynamics.

1 Introduction

The interventionist account of causation has been largely dismissed as a serious candidate for application in physics. For example, a dismissal of this sort is evident in the words of theoretical physicist Peter Havas:

We are all familiar with the everyday usage of the words “cause” and “effect”; it frequently implies the interference by an outside agent (whether human or not), the “cause”, with a system, which then experiences the “effect” of this interference. When we talk of the principle of causality in physics, however, we usually do not think of specific cause-effect relations or of deliberate intervention in a system, but in terms of theories which allow (at least in principle) the calculation of the future state of the system under consideration from data specified at a time t_0 (Havas 1974, 24).

And worries about the relevance of the interventionist account of causation in physics come not only from physicists, but also from philosophers—even those who favor interventionism:

There are important differences between, on the one hand, the [interventionist] way in which causal notions figure in common sense and the special sciences and the empirical assumptions that underlie their application and, on the other hand, the ways in which these notions figure in physics (Woodward 2007, 67).

The reasons for dismissals and worries like those above are related to a common (but problematic) assumption that causation in physics has something to do with the dynamical evolution of a closed system. The problem is that, in our preoccupation with dynamical evolution and closed systems, we tend to forget and/or neglect those areas of physics for which we do *not* have complete equations of motion or for which it *doesn't make sense* to consider entirely closed systems. And it is in those areas that the dynamical view of physical causation makes less sense and interventionism finds its home.

In this paper, I propose to take a fresh look at the interventionist account of causation and its applicability to one of those neglected areas of physics: thermodynamics. I will argue that an interventionist analysis of thermodynamics succeeds where the dynamical view of physical causation fails. As I will show, all theorizing in thermodynamics requires careful definition of the “system” under consideration, which necessarily involves attending to the boundaries that enclose the system and the conditions imposed on those boundaries. Once boundaries are adequately specified, we end up with a strong distinction between the *internal* properties and processes of the system and those *external* influences that constrain the internal dynamics. It is in the distinction between internal properties and external influences that the natural fit between the structure of thermodynamic theorizing and the interventionist account of causation becomes apparent.

The plan of this paper is as follows. In section 2, I show that interventionist reasoning is inseparable from the structural foundation of thermodynamic theory. In section 3, I show how “driving forces” and their conjugate fluxes provide a rich basis for meaningful interventionist causal claims in thermodynamics. In section 4, I use the success of interventionist causal analysis in thermodynamics to make some broader concluding remarks.

2 The centrality of manipulated equilibrium

Thermodynamic theorizing is structured around the characterization of equilibrium states and the processes by which systems move from one equilibrium state to another. But just what is a thermodynamic equilibrium state?

A thermodynamic equilibrium state is the state of a system that is *not* undergoing a change (thermal, mechanical, or chemical). However, an equilibrium state is not a spontaneous occurrence. Natural thermodynamic systems are in constant flux. They engage in all sorts of interactions: they transfer heat, push and pull on one another, change their volume, and chemically react. The very idea of a thermodynamic “system”, which can only be defined by the location and/or nature of its boundaries, is in itself a theoretical concept that we impose on the world in order to do thermodynamic “bookkeeping” (Dill and Bromberg 2011, 93). In order for a thermodynamic system to achieve an equilibrium state, the system must have been allowed to relax for a sufficient amount of time without the disturbing external influences of uncontrolled contact with other systems. And such a condition requires boundaries that isolate it—or

otherwise control exchanges—from other systems. Often those boundaries are put in place artificially, by human intervention.

Consider, for example, the air in an ordinary room. If we define our thermodynamic system in relation to the walls and doors of the room, we can say that the system has a fixed volume. If no massive weather change is currently occurring, we can assume that the air pressure in the room is approximately constant (not by isolation, but by contact with an external system whose pressure is approximately constant). If some kind of air conditioning system is in place and has been running for some time, we can also say that the temperature of the room is approximately constant. We can say that most of the chemical reactions occurring in the room are in a steady state and that the concentrations of various gases are relatively uniform (except perhaps for some minor concentration gradients near any plants and/or people located in the room), with equal flow into and out of the room for each type of gas. Notice, now, that even this *almost*-equilibrium state requires artificial maintenance (the rigidity of walls, contact with an exterior reservoir supplying constant pressure, the continuous work of the air conditioner, *etc.*). Stricter equilibrium states require much more careful isolation and maintenance, and true equilibrium states (which only exist in theory) require idealized boundaries (*e.g.*, perfect thermal insulators, frictionless pistons, perfectly rigid containers, *etc.*).

There is something of a tension, however, in the way that we think about equilibrium states. On the one hand, equilibrium states are the product of external conditions imposed on a system. On the other hand, once we consider those external conditions as given, a system will *naturally* or *spontaneously* tend toward the equilibrium state allowed by the constraints. But that spontaneous or natural behavior cannot be conceived of without external constraints being placed on the system in question. To even conceive of an equilibrium state, we must ask about the conditions imposed on its boundaries. What kind of walls enclose it? Permeable, semi-permeable, impermeable? Rigid or flexible? Adiabatic or conducting? There is no such thing as an equilibrium state unless the boundaries of the system are well-defined.¹ And the conditions imposed on those boundaries constitute external interventions on the system; they effectively *set* various thermodynamic variables to take on certain values. For example, conducting walls that put a system in contact with a thermal reservoir are effectively a way of *intervening* on temperature. Likewise, a semi-permeable boundary is a way of selectively *intervening* on particle concentrations in the system. (I will return to the question of how to conceive of boundary conditions as interventions on thermodynamic variables below in section 3.)

Thus, thermodynamic equilibrium states are inherently manipulated states—manipulated to be so either by human design or by some other mechanism that effectively imposes equilibrium conditions by external intervention. And these external manipulations or interventions, which impose values on certain thermodynamic variables, are entirely consistent with the concept of an intervention

¹In fact, a system with no defined boundaries or external constraints is effectively a universe, and its fate is something like the “heat death” discussed by Thomson, Helmholtz, and Rankine.

that has been developed by Woodward (2003) and others. According to the interventionist account of causation, an intervention directly forces a variable to take on (or remain fixed at) a certain value. Furthermore, Woodward's definition of an intervention makes no reference to human action, and thus any entity or structure playing the role of setting certain variable values or holding them fixed can fulfill the requirements for intervention. For example, a cell membrane is a structure that effectively *intervenes* to maintain a certain equilibrium internal to the cell, by keeping interior and exterior pressures equal and by maintaining certain chemical concentrations by only allowing for select passage into and out of the cell.

Now how do these manipulated equilibrium states figure into theorizing about thermodynamic processes? We begin by representing our system of interest by reference to a *thermodynamic configuration space*. The thermodynamic configuration space is the set of all possible equilibrium states of a system, where the coordinates of that space are a relatively small number of macroscopic thermodynamic variables and each point in the configuration space represents a distinct equilibrium state. For example, we might choose as coordinates the following parameters: internal energy (U), volume (V), and the particle numbers of the various species present (N_1, N_2, \dots, N_i). Then the entropy function for our system, $S = S(U, V, N_1, \dots, N_i)$, will define a hyper-surface within the configuration space (see figure 1).

With this thermodynamic configuration space and the hyper-surface defined by the entropy function in place, we can begin to theorize about any ordered sequence of states (call these A, B, C, \dots) located on the hyper-surface. Notice that a curve drawn through this sequence of states looks something like a process (in fact, we call it a *quasi-static process*) in that it represents a series of changes undergone by the system. However, such a curve can be nothing like a real process, because real processes involve nonequilibrium states and the curve represents a system that remains in equilibrium along its entire length. Furthermore, the curve could never represent the *autonomous* trajectory of a system, since every state that makes up the path is an equilibrium state and no isolated system would move from one equilibrium state to another spontaneously. So in order to think about a quasi-static process as something like a process, we must think of a system being “led”—by a series of external interventions—through the succession of desired states via “hops”. We effectively imagine the system being “corralled” through the sequence of equilibrium states. And by imagining the sequence of hops between states to be very small and carried out by very tiny interventions, we can approximate a smooth curve more and more closely (in fact, arbitrarily closely).²

In summary, the structural foundation of thermodynamic theory is the set of equilibrium states and the quasi-static “processes” that can be drawn like lines through the space of such states. As I have argued here, the very idea of an equilibrium state is not possible without reference to boundaries and the constraints that *set* the value of certain thermodynamic variables within those

²My discussion here closely follows that of Callen (1985, Ch. 4).

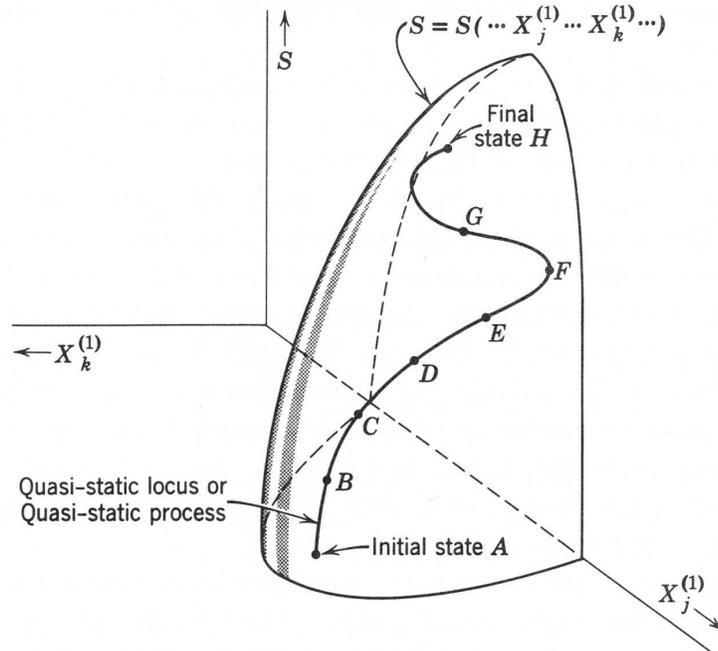


Figure 1: A representation of a quasi-static process in thermodynamic configuration space. From Callen (1985).

boundaries. Furthermore, we cannot think about quasi-static “processes”, which are sequences of those equilibrium states, without thinking about a series of infinitesimal external interventions that force a system from one equilibrium state to the next. It is in this sense that interventionist reasoning is inseparable from the structural foundation of thermodynamic theory.

In the next section, I will discuss thermodynamic theorizing in greater specificity. As I will show, the interventionist view of causation maps naturally onto the use of potential functions when theorizing about a system undergoing a process.

3 Thermodynamic potentials and driving forces

The equilibrium state toward which a system will tend, given the conditions imposed on its boundaries, is governed by the energy and entropy considerations provided in the First and Second Laws of thermodynamics. The First Law tells us that any change in the internal energy (U) of a system will be equal to the total amount of energy it gains through energy exchange with the external world, in the form of heat and/or in the form of work. The Second Law tells us that any isolated system (*i.e.*, any closed system with fixed internal energy)

will tend toward its state of maximum entropy (S). The Second Law also has the result that the internal energy of any closed system with fixed entropy will be minimized. However, neither internal energy nor entropy are directly measurable, nor do we have a specific function that tells us their dependence on other state variables. What we do have, however, are other equations of state (*e.g.*, the ideal gas law) in addition to equations for U and S in *differential* form, which tell us about the way in which small changes in other state variables relate to small changes in energy and entropy:

$$dU = TdS - pdV + \sum_j \mu_j dN_j \quad (1)$$

$$dS = \left(\frac{1}{T}\right)dU + \left(\frac{p}{T}\right)dV - \sum_j \left(\frac{\mu_j}{T}\right)dN_j, \quad (2)$$

where T is absolute temperature, p is pressure, V is volume, μ_j is the chemical potential for species j , and N_j is the number of particles for species j . The above equations (and other variant forms) are commonly referred to as *thermodynamic potential functions*.

Notice that each term in both equations above involves a pair of conjugate variables. The second term in equation 1, for example, involves pressure and volume as a conjugate pair. For every pair of conjugate variables, one of the variables is extensive (*i.e.*, additive such that the property of a system is equal to the sum of that property for all of its component subsystems), while the other is intensive (*i.e.*, independent of the size of the system). Looking again at the second term in equation 1 as an example, pressure is the intensive variable and volume is the extensive variable.

Depending on the factors controlled in a given experimental context, each pair of conjugate variables tells us something about a tendency of the system as it moves toward equilibrium in that context. Since conjugate variables will be extremely important for our purposes here, let's concentrate on one pair and use an example to decipher its practical meaning.

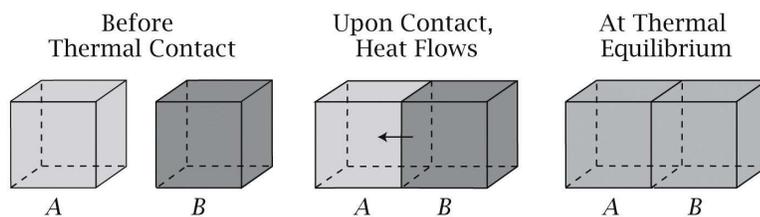


Figure 6.3 Molecular Driving Forces 2/e (© Garland Science 2011)

Figure 2: Two thermodynamic systems A and B before, during, and after arriving at thermal equilibrium. From [Dill and Bromberg \(2011, 100\)](#).

Consider the term $\left(\frac{1}{T}\right)dU$ in equation 2 and the process pictured in figure 2. We begin with two systems A and B , each enclosed in a rigid container. System A begins at temperature T_A and system B at T_B , where $T_A \neq T_B$.

The two systems are then brought into thermal contact with one another, but remain thermally insulated from the rest of the world. Now each system has an unknown entropy that can be expressed as a function of its internal energy, volume, and particle numbers, and since entropy is an extensive quantity, the total entropy of the combined system can be expressed as $S_{Total} = S_A(U_A, V_A, \mathbf{N}_A) + S_B(U_B, V_B, \mathbf{N}_B)$. Since entropy will be maximized at equilibrium, we use equation 2 to write the differential expression for S_{Total} and set it to zero:

$$\begin{aligned} dS_{Total} = & \left(\frac{1}{T_A}\right) dU_A + \left(\frac{p_A}{T_A}\right) dV_A - \sum_i \left(\frac{\mu_{A_i}}{T_A}\right) dN_{A_i} + \\ & \left(\frac{1}{T_B}\right) dU_B + \left(\frac{p_B}{T_B}\right) dV_B - \sum_j \left(\frac{\mu_{B_j}}{T_B}\right) dN_{B_j} = 0 \end{aligned} \quad (3)$$

If we assume that there is no particle exchange between the two systems and that no chemical change occurs within each system, we can eliminate the terms that allow for changing particle numbers. And since the containers are rigid, we can eliminate the terms that allow for changing volume. Furthermore, given that the combined system is isolated from the external world, the total internal energy of the combined system must remain constant, and any change in energy of either system must be compensated by a change in energy of the other. Thus, $dU_A = -dU_B$. So we have the following simplified expression:

$$dS_{Total} = \left(\frac{1}{T_A} - \frac{1}{T_B}\right) dU_A, \quad (4)$$

which will be equal to zero (*i.e.*, attain equilibrium) when $T_A = T_B$.

Thus we have derived the well-known result that two objects brought into thermal contact will reach equilibrium when their temperatures are equal. But more importantly for our purposes here, we can interpret the factors in equation 4 in light of this equilibration process. The difference in temperatures between the two systems leads to a nonzero value of the factor $\frac{1}{T_A} - \frac{1}{T_B}$, which effectively acts as a “force” driving a change dU_A in the internal energy of system A. More generally speaking, when a system is placed in thermal contact with a system at a different temperature, the temperature difference between the two systems acts as a force driving an exchange of heat energy between the systems. Phrased in terms of a system and its surroundings, $\frac{1}{T}$ describes the tendency of a system to exchange heat with its environment; it is the incremental relaxation that a system experiences in transferring a small bit of its energy dU .³

Physicists commonly use the language of “driving forces” in referring to the intensive parameters in the thermodynamic potential functions. Looking back again at equation 2, a difference between the pressure p of the system and its environment will act as a driving force for an exchange of volume dV between the system and its environment, and a difference between the concentration of a

³Alternatively, we could have begun with the thermodynamic potential function for internal energy (equation 1) to derive the same result.

particular species μ_j in the system and its environment will act as a driving force for exchanges of particles of the respective species with the environment (dN_j). The force or tendency represented in each of the conjugate pairs $(T, p, \boldsymbol{\mu})$ can act, separately or together (depending on the constraints imposed on the process), to drive changes in its paired extensive variable (dU , dV , or $d\mathbf{N}$, respectively), and thus to drive the system and its environment toward the equilibrium state of maximum entropy.⁴

This “driving force” language—and its basis in the way in which the environment exchanges energy and entropy with a system—matches the way in which relationships among thermodynamic variables would be modeled by the interventionist account of causation. According to the interventionist account, a variable X is an interventionist cause of another variable Y if there is a possible intervention on X that will change the value of Y (or the probability distribution over the values of Y) when the values of all other variables in the system remain fixed.⁵ In physical experiments, the condition that the values of all other variables in the system remain fixed across changes in the intervention on X is often enforced using what I will call “auxiliary interventions” on those variables. To see how interventionist treatment matches the “driving force” language, let’s consider the temperature equilibration case above, with system A as the causal system under investigation.

Consider the set of thermodynamic variables characterizing system A when we consider the temperature equilibration process in terms of maximization of entropy: volume V_A , the set of particle numbers for each species \mathbf{N}_A , temperature T_A , and internal energy U_A . Each of these variables is represented below in figure 3. The primary intervention in the temperature equilibration case was the operation of placing system B in thermal contact with system A . This intervention occurred specifically under conditions in which the volume V_A and particle numbers \mathbf{N}_A of system A were held constant; the enforcement of constant values of V_A and \mathbf{N}_A , by enclosing the system within rigid impermeable walls, constitutes the set of auxiliary interventions in this case. Under the conditions set by these auxiliary interventions, the primary intervention produced a change in T_A , since the original temperatures of the two systems were not equal, and this change in temperature resulted in a change in the internal energy (U_A) of the system. And since, under conditions where all other variables are held constant, the intervention was an intervention on T_A and resulted in a change

⁴Physicists use the language of “driving forces” in both the entropy and energy representations. When we flip between the energy picture of a system and the entropy picture of that same system, the metric by which we measure progress toward equilibrium changes. Each metric has its own way of characterizing the driving force because, in changing our metric of progress, there is a transformation on the force term. Still, physically, it is one and the same force driving the system toward equilibrium. This representational change in the physical equations mirrors a widely-noted feature of the interventionist account of causation: when we change the set of variables with which we characterize a causal system, our characterization of the causal relationship itself can change.

⁵I have ignored some technical details for the sake of simplicity here. See Woodward (2003, 59) for the more precise interventionist criteria for X ’s being a type-level direct cause of Y and X ’s being a type-level contributing cause of Y .

in U_A , we can say that T_A is an interventionist cause of U_A .

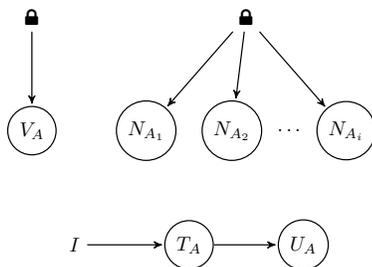


Figure 3: An interventionist causal graph of the temperature equilibration process in which system A , originally at temperature T_A , is brought into contact with another system B , originally at temperature T_B . The variable I represents the intervention that places the two systems in contact and thus changes the value of T_A . The lock symbols (🔒) represent the auxiliary interventions which hold V_A and \mathbf{N}_A fixed.

To further flesh out the causal claim being represented by the arrow from T_A to U_A in figure 3, we can contrast varying interventions in which we put system A in contact with system B at varying temperatures $T_{B1}, T_{B2}, \dots, T_{Bn}$, while still holding V_A and \mathbf{N}_A constant at the same values. Under such varying interventions, we will find that there are corresponding variations in the final T_A and U_A . Therefore, the interventionist account confirms that the temperature T_A of system A is a cause of its internal energy U_A . In general, interventions on temperature lead to changes in internal energy via exchange of heat when volume and particle numbers are held constant. Such a causal claim seems to be precisely what physicists mean to convey when they use “driving force” language with respect to temperature.

The intervention in the above case, where we have an equilibration process between two finite systems with differing initial temperatures, is an example of a “soft” or “parametric” intervention in that it *modifies* the temperature of our system rather than determining it completely.⁶ When we put system A with its initial temperature T_A in contact with system B with its initial temperature T_B , the combined system finds an equilibrium temperature somewhere between the initial values of T_A and T_B . But thermodynamics also provides conceptual tools for theorizing about “hard” or “structural” interventions that entirely determine the value of an intensive parameter for a system. We call these theoretical entities “reservoirs” or “baths”, and they have the property of being able to exchange one or more extensive quantities while their corresponding intensive properties remain constant. For example, an energy bath (*i.e.*, a temperature reservoir), by virtue of its size, is able to exchange energy with a system with which it is put in contact with negligible effect on its temperature. Likewise, a volume bath (*i.e.*, a pressure reservoir) is able to exchange volume while remaining at constant pressure, and a particle bath (*i.e.*, concentration reservoir) is able to exchange particles while maintaining constant particle con-

⁶For the distinction between soft and hard interventions, see [Eberhardt and Scheines \(2007\)](#).

centrations. When we theorize about cases in which we put a system in contact with a reservoir instead of a finite system, we consider a hard intervention that *determines* the value of the relevant intensive variable in our system. Such theoretical experiments bring the interventionist causal structure into even clearer relief: putting a system in contact with a reservoir is an intervention that sets the value of an intensive variable in the system, which in turn results in a change in the corresponding extensive variable.

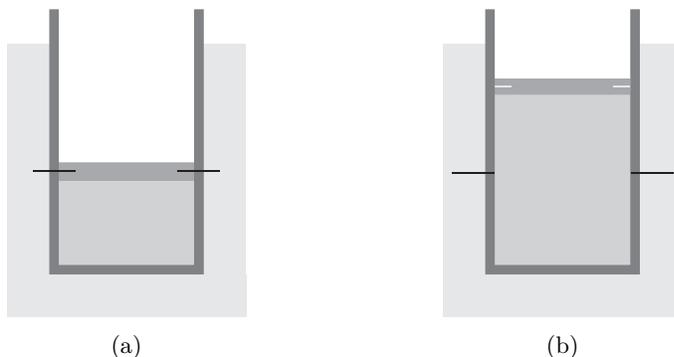


Figure 4: Illustration of a pressure-driven process, depicting (a) the system in its initial equilibrium state before the piston-locking pins are released; (b) the system once it has reached its new equilibrium state after the pins are released. This image shows the result of the case where $p_0 > p_{Res}$ and the piston rises, but all of the same considerations would apply in the case that $p_0 < p_{Res}$ and the piston falls.

Let's look at an example. Consider a system that is in an initial equilibrium state (p_0, T, \mathbf{N}) . Suppose that we intervene on the system by bringing it into contact with a reservoir that maintains the same temperature T as the system but a different pressure p_{Res} . We might do so by releasing an initially-locked piston, allowing it to move freely between the system and the reservoir (see figure 4). The process that ensues will be ruled by a maximization of the entropy of the total combined system, so we are interested in the condition where $dS_{Total} = 0$:

$$dS_{Total} = \frac{1}{T_{Res}} dU_{Sys} + \frac{p_{Sys}}{T_{Res}} dV_{Sys} + \frac{1}{T_{Res}} dU_{Res} + \frac{p_{Res}}{T_{Res}} dV_{Res} = 0 \quad (5)$$

Due to conservation of volume and conservation of energy, $dU_{Sys} = -dU_{Res}$ and $dV_{Sys} = -dV_{Res}$, so the above condition reduces to the following:

$$dS_{Total} = \left(\frac{p_{Sys} - p_{Res}}{T_{Res}} \right) dV_{Sys} = 0 \quad (6)$$

We can see here that it is the pressure difference between system and reservoir that is driving the exchange of volume. And again, this physical interpretation in terms of driving forces matches the interventionist causal account. By placing the system in contact with the reservoir, we *set* the pressure of the system to a new value, and the forced change in pressure results in a change in volume. Were we to impose a different pressure on the system by placing it in contact

with a reservoir at a different pressure, we would see the corresponding volume change as well. Thus, pressure is an interventionist cause of volume (see figure 5).

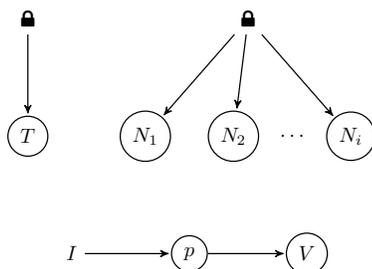


Figure 5: Interventionist causal representation of the pressure equilibration process depicted in figure 4. The variable I represents the intervention that places the system in contact with the pressure reservoir and thus changes the value of p . The lock symbols (🔒) represent the auxiliary interventions which hold T and \mathbf{N} fixed.

As shown in the examples above, the most important key to successful thermodynamic theorizing is the careful definition of the boundaries between systems and accounting for the transactions that occur at those boundaries. Interventionist reasoning fits naturally into thermodynamic theorizing because its distinction between the interventions external to a causal system and the causal relations internal to that system is perfectly applicable where thermodynamic boundaries are well-defined. Since interventions are always performed *on* a causal system from outside, it is entirely natural to label exchanges between a system and its environment as interventions of the environment on those systems.

4 Conclusion

In this paper, I have shown that there is a natural fit between thermodynamic theorizing and the interventionist account of causation. I therefore argue that the interventionist account is the most suitable account of causation for describing thermodynamic theorizing and our actual interactions with thermodynamic systems.

I suggested at the beginning of this paper that we tend to assume that physical causation will have a dynamical form, and that my identification of interventionism as the most appropriate account of causation in thermodynamics would run contrary to this assumption. It might be objected that this is a somewhat dull result, however. Thermodynamics, so the objection might run, is not “fundamental” physics, and so it is unsurprising that we should find interventionist causation rather than dynamic causation in a realm of physics that is...well...*not dynamical*. But such an objection would miss the point. Our common assumption that “physical causation” must refer to the dynamical propagations of systems is the result of our preoccupation with “fundamental” physics (which

we also assume, almost by definition, must have a dynamical form) and neglect of those areas of physics which are considered to be “non-fundamental”.⁷

So what is it to be a cause in (at least some of) physics? Here is a simple answer: an account of causation which appropriately characterizes the theoretical structure and phenomenal behavior of a domain of physics gives an account of what it is to be a cause in that domain of physics. And I have shown that the interventionist account does just that in thermodynamics.

References

- Batterman, Robert, ed. 2013. *The Oxford Handbook of Philosophy of Physics*. Oxford: Oxford University Press.
- Callen, Herbert B. 1985. *Thermodynamics and an Introduction to Thermostatistics*. 2nd ed. New York: John Wiley & Sons.
- Dill, Ken A. and Sarina Bromberg. 2011. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. 2nd ed. New York: Garland Science.
- Eberhardt, Frederick and Richard Scheines. 2007. “Interventions and Causal Inference.” *Philosophy of Science* 74 (5): 981–995.
- Havas, Peter. 1974. “Causality and Relativistic Dynamics.” In *Causality and Physical Theories*, edited by William B. Rolnick, Vol. 16 of *AIP Conference Proceedings*, 23–47. New York: American Institute of Physics.
- Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
- . 2007. “Causation with a Human Face.” Chap. 4 in *Causation, Physics, and the Constitution of Reality: Russell’s Republic Revisited*, edited by Huw Price and Richard Corry, 66–105. Oxford: Clarendon Press.

⁷Increasingly, the study of “non-fundamental” theories has revealed that their relationship with “fundamental” theories is less straightforward than might be expected. For recent discussions in this vein, see, for example, Batterman (2013). Furthermore, it is entirely unclear what the criteria for “fundamental” status are, or whether undisputed criteria even exist. And with the criteria for fundamentality in doubt, it is hard to see what basis we might have for even expecting “fundamental” theory to always have dynamical form.