PSA2020: The 27th Biennial Meeting of the Philosophy of Science Association

Baltimore, MD; 18-22 Nov 2020

Version: 3 November 2021

PhilSci ARCHIVE

PhilSci
A·R·C·H·I·V·E
philsci-archive.pitt.edu

PSA2020: The 27th Biennial Meeting of the Philosophy of Science Association
Baltimore, MD; 18-22 Nov 2020

This conference volume was automatically compiled from a collection of papers deposited in PhilSci-Archive in conjunction with PSA2020: The 27th Biennial Meeting of the Philosophy of Science Association (Baltimore, MD; 18-22 Nov 2020).

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 3 November 2021

This work is freely available online at:

http://philsci-archive.pitt.edu/view/confandvol/confandvol2020PSA.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 PSA2020: The 27th Biennial Meeting of the Philosophy of Science Association, retrieved from PhilSci-Archive at
http://philsci-archive.pitt.edu/view/confandvol/confandvol2020PSA.html,
Version of 3 November 2021, 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

Bendik Aaby, *The ecological dimension of natural selection.* 1

Mikio Akagi and Frederick W. Gooding, *Microaggressions and Objectivity: Experimental Measures and Lived Experience.* 22

Matthew J. Barker and Matthew H. Slater, *Classificatory norms in scientific practice:* 35

Andrew Bollhagen, *The "Inch-Worm Episode": Reconstituting the Phenomenon of Kinesin Motility.* 58

Julia Bursten and Catherine Kendig, *Growing Knowledge: Epistemic Objects in Agricultural Extension Work.* 85

Lorenzo Casini, Alessio Moneta, and Marco Capasso, *Variable Definition and Independent Components.* 106

David Colaço, *Believe me, I can explain! Beware of inferences to the explanandum.* 118

Devin Sanchez Curry, *g as bridge model.* 138

Andre E Curtis-Trudel, *Implementation as Resemblance.* 156

Mike Dacey, *Anecdotal Experiments: evaluating evidence with few animals.* 177

Cruz Davis, *Structural Humility.* 193

Corey Dethier, *Climate Models and the Irrelevance of Chaos.* . . . 206

Marina DiMarco, *Wishful Intelligibility, Black Boxes, and Epidemiological Explanation.* 220

Phuong (Phoebe) Dinh and David Danks, *Causal Pluralism in Philosophy: Empirical Challenges and Alternative Proposals.* 238

John Dougherty, *The substantial role of Weyl symmetry in deriving general relativity from string theory.* 258

Wei Fang, ***Towards Mechanism 2.1: A Dynamic Causal Approach.*** **269**

Luis Favela, ***"It takes two to make a thing go right": The coevolution of technological and mathematical tools in neuroscience.*** **294**

Enno Fischer, ***Causation and the Problem of Disagreement.*** **317**

Justin Garson, ***Edmond Goblot's (1858-1935) Selected Effects Theory of Function: A Reappraisal.*** **332**

Christopher Grimsley, ***Causal and Non-Causal Explanations of Artificial*** **345**

Stephanie Harvard, ***Representational Risk.*** **365**

Josh Hunt, ***Understanding and Equivalent Reformulations.*** **366**

Kati Kish Bar-On, ***How Much Change is Too Much Change? Rethinking the Reasons Behind the Lack of Reception to Brouwer's Intuitionism.*** **378**

Travis LaCroix, ***Reflexivity, Functional Reference, and Modularity: Alternative Targets for Language Origins.*** **401**

Matthew J. Maxwell, ***The Evidence-Observation Distinction in Observation Selection Effects.*** **415**

Michael Miller, ***Infrared Cancellation and Measurement.*** **435**

Vitaly Pronskikh, ***Engineering roles and identities in the scientific community: toward participatory justice.*** **449**

Samuli Reijula and Jaakko Kuorikoski, ***The diversity-ability trade-off in scientific problem solving.*** **460**

Sarita Rosenstock, ***Learning from the Shape of Data.*** **480**

Ameer Sarwar, ***Perspectives on Causal Specificity.*** **495**

Mike D. Schneider, ***Creativity in the social epistemology of science.*** **512**

Gerhard Schurz, ***Tacking by Conjunction, Genuine Confirmation and Bayesian Convergence.*** **531**

Michael Silberstein and W. M. Stuckey, *Beyond Causal Explanation: Einstein's Principle Not* 547

W. M. Stuckey, Michael Silberstein, and Timothy McDevitt, *A Principle Explanation of Bell State Entanglement.* 576

Morgan Thompson, *Epistemic Risk in the Triangulation Argument for Implicit Attitudes.* 599

Aja Watkins, *The Epistemic Value of the Living Fossils Concept.* . 616

Tung-Ying Wu, *Structural Decision Theory.* 639

Carlos Zednik and Hannes Boelsen, *The Exploratory Role of Explainable Artificial Intelligence.* 655

The Ecological Dimension of Natural Selection

Abstract: In this paper I argue that we should pay extra attention to the ecological dimension of natural selection. By this I mean that we should view natural selection primarily as acting on the outcomes of the interactions organisms have with their environment which influences their relative reproductive output. A consequence of this view is that natural selection is not (directly) sensitive to what system of inheritance which ensures reoccurrences of organism-environment interactions over generations. I end by showing the consequences of this view when looking at how processes like niche construction and the Baldwin effect relate to natural selection.

1. Introduction. The principle of natural selection is the theoretical cornerstone of evolutionary theory. In the philosophy of biology, we can delineate four different, but related, main discussions of this principle; first, on what the sufficient conditions are for its occurrence (e.g., Lewontin 1970; Godfrey-Smith 2009). Second, on the appropriate means of quantifying the influence of natural selection on the distribution of variants in populations over time (e.g., Millstein 2009; Otsuka 2016). Third, on whether selection can be counted as a cause or is more appropriately interpreted as a statistical summary of multiple underlying causes and not a cause of evolution in itself (e.g., Matthen and Ariew 2002; Ramsey 2013ab; Walsh 2010). Fourth, on whether selection can act on multiple levels and what the relevant units of selection are, and if any of these are privileged (e.g., Williams 1966; Dawkins 1976; Okasha 2006).

Another debate, which is related to all of the aforementioned debates, centers around the metaphysics of evolution. In this debate we can identify two main camps; a molecular, or "gene-centered" metaphysics (e.g., Dawkins 1976, 1982) and an ecological, or "organism-centered" metaphysics (e.g., West-Eberhard 2003; Walsh 2015). Standard textbook evolutionary biology usually has a "molecular" metaphysics, in that the fundamental units of evolution are *genes*. On an "ecological" metaphysics of evolution, the fundamental unit of evolution are *organisms*.

Walsh (2015), amongst others (see references below), has recently argued that the Modern Synthesis misrepresents the metaphysics of evolution by viewing it primarily as a molecular phenomenon, instead of an ecological one. This is largely due to what Walsh

calls “the marginalisation of the organism that have taken hold under the Modern Synthesis” (Walsh 2015, *x*). This has been a complaint of many biologists and philosophers over the last decades (e.g., Lewontin 1983, Piaget 1978; Odling-Smee et al. 2003; Oyama 2000; West-Eberhard 2003) and is a central complaint of the proponents of an *extended evolutionary synthesis* (Pigliucci and Müller 2010). Theoretical and empirical work taking a more ecological or organism-centered approach to understanding evolution and development has also recently gained some traction under the headings of eco-devo (ecological developmental biology) and eco-evo-devo (ecological evolutionary developmental biology). For example, West-Eberhard (2003), Sultan (2015) and Gilbert and Epel (2015) have made a great effort to establish how both evolutionary and developmental trajectories are significantly influenced by, and sometimes crucially dependent on, particular organism-environment interactions.

This paper is a philosophical contribution to what an “organism-centered”, or “ecological”, metaphysics of evolution might do to our understanding of natural selection. I begin from the view that natural selection is primarily an ecological process. By this I mean that natural selection is a process where organism-environment interactions are what is preferentially selected. Further, natural selection acts on the outcomes of these interactions. This is not a novel view and has been suggested before (Lehrman 1970; Brandon 1990; Rosenberg 1983). However, I will take this a step further and argue that this also means that natural selection is not directly sensitive to which system of inheritance ensures the reoccurrence of such interactions, be it genetic, epigenetic, behavioral, cultural, or symbolic (Jablonka and Lamb 2014). Natural selection acts on the

outcomes of organism-environment interactions and the frequency and likelihood of their reoccurrence in subsequent generations.

However, this does not mean that I equate the importance of each system of inheritance. A genetic system of inheritance is an important prior condition for there to be other systems of inheritance in most, if not all, organisms. Further, most of morphological and physiological evolution seem to be primarily under genetic control. The point is rather that this happens "unbeknownst" to natural selection. To use some helpful terminology from Sober (1984), we can say that there is selection *for* the ecological interactions that yields highest relative fitness in a population, while there is selection *of* the relevant genes that contribute to those interactions because of the high-fidelity-inheritance properties of the genetic system of inheritance in reliably producing offspring which have similar interactions.

2. Selection on Passive Objects by Environmental Filtration. Let us begin by looking in more detail at the "standard" molecular metaphysics of the Modern Synthesis. In most textbooks on evolutionary biology, one is likely to find a definition of evolution as the changes to allele (or gene) frequencies in a population over time (e.g., Futuyma and Kickpatrick 2017). Furthermore, the conditions for evolution by natural selection to occur (e.g., Lewontin 1970); inheritance, variation, and differences in fitness, is often interpreted in a genetic manner. That is, any variation in fitness, which is due to differences in the performance of varying phenotypes in relation to the local (and shared) selective environment, is only acted on by natural selection insofar as the genetic underpinning of

that variation steadily expresses the relevant phenotype over generations. Since the genetic system of inheritance is privileged, in the sense that without it there would be (in most cases) no organism to be selected for in the first place, it makes perfect sense to define evolution as changes in the frequencies of genes in a population. And from this it easy to conceive of natural selection as being an agent which sorts different genetic variants based on their performance relative to their immediate environment. This rendition of natural selection construes it as an environmental process. The metaphor of a sieve or filtration is often invoked to describe this process (e.g., Sober 1984). Coupled with the view that the only phenotypic variation that matters for biological evolution is that which is the result of genetic variation, such metaphors engender a certain passivity on behalf of the organism. It essentially relegates the action of selection to be realized by certain (stable or changing) environmental configurations. Natural selection acts on those organisms that carry the appropriate genetic material to produce a phenotype that performs best (i.e., highest realized relative fitness) in relation to the relevant environmental configurations. Such a view of evolution by natural selection has been called asymmetrically externalist (Godfrey-Smith 1996). It is asymmetric in the sense that the configurations of the environment are (presumed to be) explainable solely with reference to factors internal to the environmental system itself. While, on the other hand, the organisms which occupy these environments are explained (in terms of the phylogenetic history leading up to their capacity for occupying the environment) by reference to a combination of changes to the biological system (i.e., changes in the gene frequencies of the lineage(s) leading up to the relevant population) and the environmental configuration which the lineage(s) have experienced

over generations. It is externalist in the sense that the environmental configurations are what "trigger" the selection of the phenotype, while the changes to the gene frequencies in the population is a "structuring" cause of the selection event.[1] The role of the organism in such explanations is that of a vehicle (Dawkins 1978), one that carries certain passengers (genes) to certain destinations (selection events). However, organisms are arguably not just an ensemble of genes, and their activity or behavior might influence their reproductive success and consequently the evolution of their lineage. How does an externalist and molecular (i.e., gene-centered) view of evolution deal with behavior?

Standardly, in behavioral ecology (e.g., Krebs and Davies 1993) and the evolutionary explanations provided by behavioral genetics (e.g., Anholt and Mackay 2010), organismic activity and behavior is treated as any other phenotypic trait. It is based on certain assumptions regarding the dispositional properties of genes in relation to behaviors and certain optimality measures (Krebs and Davies 1993). Generally speaking, organisms exhibiting behaviors that increase their fitness are selected for, and the disposition to exhibit the beneficial behavior in subsequent generations is assumed to be under genetic control—and can consequently be treated like any other phenotypic trait. The validity of these assumptions is not under question here. The point here is a conceptual one. It is about how we conceive of the relation between natural selection and the organisms exhibiting the relevant behavior. Let us do a thought experiment. Take an

[1] For the distinction between "structuring" and "triggering" causes, see Dretske (1988). For an example of its relevance for evolutionary theory, see Ramsey (2016).

imaginary species like the *tarbutniks* from Avital and Jablonka (2000). The individuals of this species have completely identical and non-changing genetic make-up. In other words, it is a species without genetic variation among the individuals. However, let us assume that they can differ in their behavior, i.e. that there is still phenotypic variation. Let us then imagine that some individuals forage fruits to supplement their diet, while others obtain their nutrients from only eating grass. This then leads to the fruit-foraging individuals having a more energy-rich diet, which increases their reproductive output. Let us further imagine that the fruit foraging techniques are passed on vertically through parental guidance (i.e., learning) and that the transmission of this behavior from parent to offspring enjoys a high level of fidelity. If we view natural selection as a process that sorts genetic variation, then there is no response to selection in this scenario. However, this seems wrong. Surely, natural selection still acts on the individuals that forage fruit to supplement their diet if this increases their reproductive output. Thus, there is a response to selection in the population—the number of fruit-foraging individuals increases and fruit-foraging behavior spreads throughout the population.

While in this thought experiment natural selection does not lead to biological evolution (in the sense that the gene frequencies in the population remain unchanged), natural selection has still occurred. And while it might be true that for natural selection to bring about adaptive *biological* evolution there must selection amongst different genetic variants in a population, there is still natural selection amongst the phenotypes of our imagined population. The strength and direction of the selection for the fruit foraging behavior is dependent on the fidelity and transience of the behavioral inheritance system.

Even though there are no organisms like the *tarbutniks* in the real world and we do not know exactly to what extent difference in behavior and capacity for learning is linked to and/or governed by genetic variation in a population, the point about the natural selection being an ecological process still stands. Natural selection is not directly sensitive to what causes the phenotypic variation available for selection to act on, just the outcome of different interactions between phenotypes and their environments. This is an important consideration for both biologists and philosophers taking a more organism-centered approach. These argue that organisms are not merely passive objects of selection, but active subjects—or agents—in their own evolution (e.g., Lewontin 1983; Odling-Smee 2003; Bateson 2004). Let us now turn to these organism-centered views, and in particular two processes where the activities of organisms play an important part in shaping evolutionary dynamics—the Baldwin effect and niche construction.

3. Organisms as Agents in Evolutionary Theory. Over the course of the last decades there has been an increasing tension in evolutionary biology, culminating in an overarching debate surrounding whether an extended evolutionary synthesis is needed (Müller and Pigliucci 2010, Laland et al. 2014; Wray et al. 2014). A central part of this debate concerns the role that behavior, and organismic activity more generally, has on evolutionary dynamics. The question of how the activities and behaviors of organisms can alter the action of natural selection has a long history. It could, arguably, be said to date all the way back to Lamarck (Avital and Jablonka 2000). Alternatively, we can trace it back to the introduction of organic selection (also called the Baldwin effect) in the late 19th century

(Baldwin 1896a, 1896b; Morgan 1896; Osborn 1896). Organic selection refers to an evolutionary process that can turn acquired characters into congenital ones. More precisely, it refers to a three-step process; first, organisms can through their interactions with the environment systematically produce behavioral, morphological, or physiological modifications that are not hereditary, but increase the fitness of the organism that acquires them. Second, there is genetic variation in the population producing hereditary characters similar to characters that are acquired by the organisms through their environmental interactions. Third, this genetic variation is acted on by natural selection and subsequently spread in the population over the course of generations. The character was initially individually *acquired*, but is in time turned into a *hereditary* character (Simpson 1953). This process has recently garnered more attention in evolutionary biology. In the works of the late Patrick Bateson (2004, 2017a, 2017b; Bateson and Gluckman 2011) this process is revisited in light of what we have learned about social learning, transmission and non-genetic systems of inheritance over the last decades. Bateson refers to the Baldwin effect as the *adaptability driver* (Bateson 2017a). By this he means that, more often than what we initially may have thought, behavioral plasticity (behavior which is the result of stimuli or interactions with the environment, and not determined by genetic factors) is actually crucial in initiating adaptive responses to environmental challenges.[2]

[2] A more general rendition of this view, where not only behavioral but also morphological and physiological acquired characters are what initiates evolutionary change, is referred to as 'plasticity-first evolution' (e.g. Levis and Pfennig 2016).

Another example of organismic activity altering evolutionary dynamics can be seen in niche construction theory (Odling-Smee 2003). Niche construction refers to cases where organisms modify selection pressures by actively altering their environment or their relationship to it. The paradigmatic example being the beaver, which significantly alters the local environment by building a dam, and consequently altering the selective environment it experiences. Both the Baldwin effect and niche construction are central elements in the discussion of an extended evolutionary synthesis. The argument for an extended synthesis from niche construction theory is that viewing organisms as merely passive objects that are filtered by natural selection neglects the active role of the organism in its evolution (Odling-Smee 2003). They see niche construction as an evolutionary process whereby the activities of organisms counter or direct the action of natural selection. Consequently, they argue that niche construction should be seen as a potentially equally important evolutionary process as natural selection itself. The same is often said of the Baldwin effect. It constitutes a corollary process of selection (viz. organic selection) and is often considered to be an evolutionary mechanism or process (Bateson 2017a, 2017b).

According to the adherents of an extended evolutionary synthesis, we need to pay more attention to the neglected process of niche construction, organic selection and other processes where organisms play an active role in evolution. Allowing more processes to be considered *evolutionary* processes is one way we can do this (Scott-Phillips et al. 2014; Laland 2015). However, this solution has been met with some skepticism (e.g. Welch 2017; Scott-Phillips et al. 2014), as it is unclear whether granting something the status of

an evolutionary process actually increases our understanding of evolution. Another problem with viewing niche construction as an evolutionary process that counteracts natural selection is that it still treats natural selection as an asymmetrically externalist environmental process. If niche construction "counteracts" the action of selection, selection must be a process that runs from the environment to the organism. Instead, we should start from an ecological metaphysics of evolution (Walsh 2015).

4. An Ecological Metaphysics of Evolution and Organism-Environment Interactions.

When Walsh (2015) calls for an ecological metaphysics of evolution, he highlights that we might have missed a lot in our understanding of evolution by not seeing organisms as active (and purposive) agents in their environments. Treating organisms as biological agents prior to evolutionary agents is a necessary step in the direction of an ecological metaphysics (Walsh 2015). Biological entities are entities that interact with their environment. The relationship between the organism and the environment is crucial and in a sense prior to both the organism and environment themselves. Without any organisms there would be no environments, and conversely, without environments there would be no organisms (Lewontin 2000). From an ecological metaphysics of evolution, then, the fundamental unit is that of organism-environment interactions. Evolution concerns changes in the types of interactions there are. Mostly these interactions change in virtue of changes to the organism itself, for example by organism evolving faculties with which they interact with their environment in novel ways. Such kinds of changes to organism-environment interaction are captured by the theoretical framework offered by the modern synthesis.

However, an environment can also change in such a way that organism-environment interactions change as a result, and more importantly, an organism can change the environment or its relationship to it such that the organism-environment interactions change (i.e., niche construction).

Natural selection, then, is the process whereby organism-environment interactions are preferentially selected. It is concerned with the outcomes of organism-environment interactions over the life-history of an organism (or at least to the end of its reproductive age) relative to those of its population. The strength of and response to selection is determined by the probability that advantageous interactions reoccur in subsequent generations. Consequently, advantageous hereditary traits (traits that are passed on through genetic inheritance) are more likely to spread than acquired traits whose likelihood of reoccurrence is lower. But it is in principle possible for selection to act on advantageous organism-environment interactions that are acquired (e.g., as a result of niche construction or behavioral plasticity).

Take, for instance, gastrolith usage. Gastroliths are small stones that are ingested and then reside in the gastrointestinal tract of some animals. Carrying gastroliths is certainly an example of an acquired trait, as it is something the animal has to *acquire* from its environment to utilize. Usage of gastroliths is quite common among some groups of vertebrates and may serve a wide variety of different functions in relation to different environments (Wings 2007). For example, some have argued that in aquatic environments gastroliths might be used as ballast or for buoyancy control (Rondeau et al. 2005). While in

terrestrial environments some have argued that gastroliths may supply minerals and help with trituration and mixing of foodstuffs (Wings 2007).

If, for instance, an organism enjoys a higher fitness relative to other members of its population as a result of having ingested gastroliths, natural selection will favor that individual. Further, let us say that this organism learnt to ingest gastroliths by observing its parents and continue the habit of ingesting such stones. If in the subsequent generations gastrolith ingestion is reliably transmitted through observational learning, and the fitness advantage is sufficiently high, natural selection could spread this trait throughout the population. Natural selection could also favor those who have a disposition for ingesting gastroliths, with or without observational learning, making it an acquired trait with a hereditary basis (which is an example of the Baldwin effect). For natural selection, however, the basis on which the gastrolith is ingested—be it by way of learning or instinct—is irrelevant as long as the stone is ingested. It is the outcome of the interaction—e.g., the improved trituration of foodstuffs—which is conducive to the fitness advantage, not whether or not it is learnt or instinctual, as long as the stone is reliably ingested.[3] More generally, we could say that the primary way in which genes matter for selection is in how conducive they are to the reliability and likelihood of advantageous organism-environment

[3] Of course, if *all* members of a population ingest gastroliths, and some do it instinctually while others need to learn it through observation, natural selection will most likely favor the instinctual response because the trait itself (i.e., gastrolith ingestion) is presumably transmitted with a higher fidelity if it is congenital rather than learned.

interactions to reoccur in subsequent generations. Taking this perspective on how natural selection acts, let us return to how we should interpret niche construction and the Baldwin effect. Are they different selective processes, as it is commonly argued?

5. Niche Construction and the Baldwin Effect Revisited. Both niche construction and the Baldwin effect have been seen as distinct evolutionary mechanisms or processes (e.g., Odling-Smee et al. 2003; Bateson 2017a, 2017b). Some even go as far as saying that they are different *selective* processes, as when niche construction is interpreted as a process where organisms counteract natural selection by modifying selection pressures (Laland 2015). The Baldwin effect is seen as a distinct selective process which operates on acquired traits until there is genetic variation present so natural selection can "take over" and consequently turn them into congenital traits.

I think these interpretations are misguided, and stem from viewing natural selection as a process of environmental filtration concerned with primarily with genes, i.e., from a "molecular" metaphysics of evolution. If we instead take the point of view introduced above, where natural selection is concerned with the outcomes of organism-environment interactions and their relative reoccurrence, niche construction and the Baldwin effect are ways in which adaptation can occur and consequently be selected for. Niche construction is one way in which an organism can achieve a fitness advantage relative to other members of its population, but it is not a process that is counteracting the effects of natural selection. As long as the niche constructing behavior reoccurs reliably and the altered ecological conditions are reliably transmitted across generations it is no different from any other

phenotypic trait in relation to natural selection. Acquired traits, and the Baldwin effect more generally, are also not selected initially by a process distinct from natural selection (i.e., organic selection). They are selected for by natural selection from their first occurrence, it is just a shift in the system of inheritance that is responsible for the reoccurrence of the trait. Sometimes, it makes sense to say that an acquired trait has become a congenital trait, as for instance when a learnt behavior has become instinctual. However, in the case discussed above, the ingestion of gastroliths, it is unclear if it can ever fully be a congenital trait, as the key feature of having that trait is to *acquire* a suitable rock from the environment (though the disposition can certainly be congenital).

Natural selection understood as a process acting on the outcome of reoccurring organism-environment interactions has the benefit of being compatible with the main insights of the modern synthesis, while also allowing for other cases to be included as ways in which organism-environment interactions can change and be acted on by selection, such as niche construction and the Baldwin effect. It explains why the genetic system of inheritance is so important—because it is a system which is necessary for the development of (most, if not all) phenotypes, and consequently for there to be any organism-environment interactions at all. While simultaneously explaining how certain behavioral innovations, cultural traits, etc. can be selected for by natural selection, without being (directly) dependent on genetic variation or inheritance.

6. Conclusion. I have argued that natural selection is standardly understood as a process of environmental filtration concerned primarily with genes. Further, I followed Walsh (2015)

in arguing that this stems from a gene-centered and externalist ("molecular") metaphysics of evolution. If we instead opt for an ecological metaphysics of evolution our understanding of natural selection becomes different. On such a metaphysics of evolution, natural selection becomes a process that acts on the outcomes of the advantageous interactions an organism has with its environment during its life-history. As long as such interactions reoccur reliably in subsequent generations, natural selection will be insensitive as to what brings about these interactions, be it through genetic inheritance, social learning, cultural transmission, etc. A benefit of this view is that the ecological account of natural selection is compatible with the main insights from the modern synthesis, while also allowing for phenomena traditionally excluded from the modern synthesis, but emphasized by the extended evolutionary synthesis. Finally, the ecological view of natural selection can integrate some of these novel phenomena easily, without having to supplement and extend evolutionary theory with a host of new evolutionary processes.

REFERENCES

Anholt, R. and Mackay, T. 2010. *Principles of Behavioral Genetics.* London: Elsevier Academic Press.

Avital, E. and Jablonka, E. 2000. *Animal Traditions: Behavioural Inheritance in Evolution*. Cambridge: Cambridge University Press.

Baldwin, J. M. 1896a. "A New Factor in Evolution." *American Naturalist*. 30: 441–451.

Baldwin, J. M. 1896b. "A New Factor in Evolution (Continued)." *American Naturalist*. 30: 536–553.

Bateson, P., 2004. "The Active Role of Behaviour in Evolution." *Biology and Philosophy*. 19: 283–298.

Bateson, P., 2017a. "Adaptability and Evolution." *Interface Focus* **7.**

Bateson, P., 2017b. *Behaviour, Development and Evolution*. Cambridge: Open Book Publishers.

Bateson, P. and Gluckman, P. 2011. *Plasticity, Robustness, Development and Evolution*. Cambridge: Cambridge University Press.

Brandon, R. N. 1990. *Adaptation and Environment*. Princeton, NJ: Princeton University Press.

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Dawkins, R. 1978. "Replicator Selection and the Extended Phenotype." *Ethology*. 47(1): 61–76.

Dawkins, R. 1982. *The Extended Phenotype*. Oxford: Oxford University Press.

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

Futuyma, D. J. and Kirkpatrick, M., 2017. *Evolution*. Sunderland, MA: Sinauer Associates Inc.

Gilbert, S. F. and Epel, D. 2015. *Ecological Developmental Biological: The Environmental Regulation of Development, Health, and Evolution.* Sunderland, MA: Sinauer Associates Inc.

Godfrey-Smith, P. 1996. *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.

Godfrey-Smith, P. 2009. *Darwinian Populations and Natural Selection*. Oxford: Oxford University Press.

Jablonka, E. and Lamb, M. J. 2014. *Evolution in Four Dimension: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge, MA: The MIT Press.

Krebs, J. R. and Davies, N. B. 1993. *An Introduction to Behavioural Ecology*. Malden, MA Blackwell Science Ltd.

Laland, K. N., 2015. "On Evolutionary Causes and Evolutionary Processes." *Behavioral Processes*. 117: 97–104.

Laland, K. N., Feldman, M. W., Müller, G. B., Jablonka, E. Uller, T., Sterelny, K., Moczek, A., Odling-Smee, J., 2014. "Does Evolutionary Theory Need a Rethink? Yes, Urgently." *Nature*. 514(7521): 161–164.

Lehrman, D. S. (1970). "Semantic and Conceptual Issues in the Nature-Nurture Problem." In L. R. Aronson, E. Tobach, D. S. Lehrman, and J. S. Rosenblatt (eds.), *Development and Evolution of Behavior: Essays in Memory of T. C. Schneirla*. , San Francisco: W. H. Freeman, pp. 17–52.

Levis, N. A. and Pfennig, D. W. 2016. "Evaluating 'Plasticity First Evolution in Nature: Key Criteria and Empirical Approaches." *Trends in Ecology and Evolution*. 31(7): 563–574.

Lewontin, R. C. 1970. “The Units of Selection.” *Annual Review of Ecology and Systematics*. 1: 1–18.

Lewontin, R. C. 1983. “The Organisms as the Subject and Object of Evolution.” *Scientia*. 118: 63–82.

Lewontin, R. C. 2000. *The Triple Helix: Gene, Organism, and Environment*. Cambridge, MA: Harvard University Press.

Matthen, M. and Ariew, A. 2002. “Two Ways of Thinking about Fitness and Natural Selection. *Journal of Philosophy*. 99(2): 55–83.

Millstein, R. L. 2009. “Populations as Individuals.” *Biological Theory*. 4(3): 267–273.

Morgan, C. L. 1896. *Habitat and Instinct*. New York, NY: Edward Arnold Press.

Odling-Smee, J., Laland, K. N., and Feldman, M. W. 2003. *Niche Construction: The Neglected Process in Evolutionary Theory*. Princeton, NJ: Princeton University Press.

Okasha, S. 2006. *Evolution and the Levels of Selection*. Oxford: Oxford University Press.

Osborn, H. F. 1896. “A Mode of Evolution Requiring Neither Natural Selection nor the Inheritance of Acquired Characters.” *Transactions of the New York Academy of Sciences*. 15: 141–142.

Otsuka, J. 2016. “Causal Foundations of Evolutionary Genetics.” *The British Journal for the Philosophy of Science*. 67(1): 247–269.

Oyama, S. 2000. *The Ontogeny of Information: Developmental Systems and Evolution*. Durham, NC: Duke University Press.

Piaget, J. 1978. *Behavior and Evolution*. New York: Pantheon Books.

Pigliucci, M. and Müller, G. B. 2010. *Evolution: The Extended Synthesis*. Cambridge, MA: The MIT Press.

Ramsey, G. 2013a. "Can Fitness Differences Be a Cause of Evolution?" *Philosophy and Theory in Biology*. 5:e401.

Ramsey, G. 2013b. "Organisms, Traits, and Population Subdivisions: Two Arguments Against the Causal Conception of Fitness?" *British Journal for the Philosophy of Science*. 64: 589–608.

Ramsey, G. 2016. "The Causal Structure of Evolutionary Theory." *Australasian Journal of Philosophy*. 94: 421–434.

Rondeau, S. L. and Gee, J. H. 2005. "Larval Anurans Adjust Buoyancy in Response to Substrate Ingestion." *Copeia*. 2005(1): 188–195.

Rosenberg, A. 1983. "Fitness." *Journal of Philosophy*. 80(8): 457–473.

Scott-Phillips, T. C., Laland, K. N., Shuker, D. M., Dickins, T. E., and West, S. A., 2014. "The Niche Construction Perspective: A Critical Appraisal." *Evolution*. 68(5): 1231–1243.

Simpson, G. C. 1953. "The Baldwin Effect." *Evolution*. 7: 110–117.

Sober, E. 1984. *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. Chicago, IL: The University of Chicago Press.

Sultan, S. E. 2015. *Organism & Environment: Ecological Development, Niche Construction, and Adaptation*. Oxford: Oxford University Press.

Walsh, D. M. 2010. "Not a Sure Thing: Fitness Probability, and Causation." *Philosophy of Science*. 77(2): 147–171.

Walsh, D. M. 2015. *Organisms, Agency, and Evolution*. Cambridge: Cambridge University Press.

Welch, J. J. 2017. "What's Wrong with Evolutionary Biology?" *Biology and Philosophy*. 32(2): 263–279.

West-Eberhard, M. J. 2003. *Developmental Plasticity and Evolution*. Oxford: Oxford University Press.

Williams, G. C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, NJ: Princeton University Press.

Wings, O. 2007. "A Review of Gastrolith Function with Implications for Fossil Vertebrates and a Revised Classification." *Acta Palaeontologica Polonica*. 52(1): 1–16.

Wray, G. A., Hoekstra, H. E., Futuyma, D. J., Lenski, R. E., Mackay, T. F. C., Schluter, D. and Strassmann, J. E. 2014. "Does Evolutionary Theory Need a Rethink? No, All is Well." *Nature*. 514: 161–164.

Microaggressions and Objectivity: Experimental Measures and Lived Experience

Mikio Akagi[†‡]
John V. Roach Honors College, Texas Christian University
Frederick W. Gooding, Jr.
John V. Roach Honors College, Texas Christian University

Microaggressions are, roughly, acts or states of affairs that express prejudice or neglect toward oppressed group members in relatively subtle ways. There is an apparent consensus among both proponents and critics of the MICROAGGRESSION concept that microaggressions are "subjective." We examine what subjectivity amounts to in this context and argue against this consensus. We distinguish between microaggressions as an explanatory posit and microaggressions as a hermeneutical tool, arguing that in either case there is no reason at present to regard microaggressions as subjective, and that microaggressions in the hermeneutical sense should be regarded as objective.

1. Introduction. The MICROAGGRESSION concept has received much attention—both scholarly and popular—over the last decade. Microaggressions are, roughly, acts (often but not exclusively speech acts) that exhibit prejudice or neglect toward members

[†] To contact the authors, please write to: Texas Christian University, TCU Box 297922, Fort Worth, Texas 76129; email: m.akagi@tcu.edu.

[‡] This paper is the result of ongoing discussions between the authors; Mikio Akagi is responsible for much of the specific language in the paper; Frederick Gooding contributed to many stages of planning, discussion, and revision. We thank, for their input and encouragement: Karen Kovaka, Nicholas Zautra, Rob Garnett, Andrew Ryder, and participants at the inaugural conference of the Mid/South Philosophy of Science Network and the TCU Interdisciplinary Works in Progress talk series.

This is a preprint. Final version to be published in Philosophy of Science.

of oppressed groups, or states of affairs that exclude or denigrate members of oppressed groups. There appears to be a consensus that microaggressions are "subjective," both among the concept's scientific proponents, such as Derald Wing Sue, and its critics, such as Scott O. Lilienfeld. Presumably, the claim that microaggressions are "subjective" means that there is no perspective-independent matter of fact regarding whether an act or state of affairs is a microaggression. That is, whether an act or state of affairs counts as a microaggression depends upon how it is perceived by some subject. We disagree with this consensus, distinguishing between "explanatory" and "hermeneutical" MICROAGGRESSION concepts. We argue that there is no a priori reason to regard explanatory microaggressions as "subjective," and that there are compelling phenomenological reasons to regard hermeneutical microaggressions as objective.

2. Microaggressions and their effects. The term "microaggression" was coined by African American psychiatrist Chester Pierce (1970) as a label for subtle forms of hostility or disdain commonly exhibited by White Americans against African Americans. The term was subsequently amplified by psychologist Derald Wing Sue and colleagues (Sue et al. 2007), who generalized the concept to encompass many subtle forms of racism. Their oft quoted gloss is that:

> Racial microaggressions are brief and commonplace daily verbal, behavioral, and environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults to the target person or group. (Sue et al. 2007, 273)

The term is now understood broadly, both in critical theory and in psychology, as including not only racial slights, but also those related to gender (Capodilupo et al. 2010; Barthelemy et al. 2016), LGBTQ oppression (Nadal, Rivera, et al. 2010), disability (Keller and Galgay 2010; Gonzalez et al. 2015), socioeconomic status (Smith and Redington 2010), religion (Nadal, Issa, et al. 2010), or indeed any form of structural oppression (Sue 2010c), including intersectional forms of oppression (q.v. Crenshaw 1989; Nadal et al. 2015; Olkin et al. 2019). A person or social group that is demeaned or alienated by a microaggression is called a *target*. For microaggressions that are acts, the agent of the microaggression is generally called a *perpetrator* or *performer*.

A commonly-cited example of a verbal microaggression (e.g. in Sue 2010a; Lilienfeld 2017) is a remark made by John McCain during his 2008 presidential campaign against Barack Obama. A woman at a town hall event said to McCain that

she doesn't trust Obama because "He's an Arab." McCain replied, "No ma'am. He's a decent family man, citizen, that I just happen to have disagreements with on fundamental issues… He's not." McCain's reply carries the unfortunate (and probably unintentional) conversational implicature that being of Arab descent counts in some way against being "a decent family man" or a "citizen." As such, it is an ethnic microaggression. Many microaggressions carry such implicatures, which are referred to in the microaggression literature as *hidden messages* (Sue et al. 2007). One of the challenges for researchers who generalize the MICROAGGRESSION concept to new domains of oppression is the identification of the relevant hidden messages (Johnston and Nadal 2010). Microaggressions can also be nonverbal acts, e.g. tightly clutching one's purse or crossing the street when encountering a Black man. And microaggressions can be states of affairs, such as the persistence of a problematic monument. Sue and colleagues (2007) call these latter states of affairs *environmental microaggressions*.

Some (e.g. Greg Lukianoff and Jonathan Haidt and other critics of "campus culture") suggest that the proper response to microaggressions is to toughen up or "grow a thicker skin." As Regina Rini notes, this may be an appropriate response to mere insults but it is an insufficient response to microaggressions because microaggressions are components of larger patterns of systematic oppression (2018). The targets of microaggressions are necessarily oppressed groups or their members. Of course slights can target privileged social groups or their members (e.g. "White people can't dance"), but such slights are not called "microaggressions" because they are not likely to have the same negative effects.[1] The relevant difference between microaggressions and other slights is that microaggressions are *congruent* with oppressive systems, in Liao and Huebner's (2020, 10) sense, and therefore are smaller extensions of larger power structures. Slights that target privileged social groups go against the grain of oppressive social systems rather than being congruent with them.

Rini's reply is underappreciated in many skeptical discussions of microaggressions, including Lilienfeld's (2017), which raises doubts about whether the acts called microaggressions are always performed with malicious motivations. Performers' motivations may be relevant for assigning blame (see Washington and Kelly 2016 for

[1] We recognize standard provisos here: individual persons can be members both of oppressed and privileged social groups; e.g. a wealthy queer person may experience structural disadvantage related to their queerness but privilege related to their socioeconomic class. And oppression often compounds in a non-additive manner for those who are members of multiple oppressed social groups, e.g. Black women in the U.S. experience specific challenges faced neither by Black men nor by White women (Crenshaw 1989).

discussion), but not for understanding the effects of microaggressions on their targets. Much of the psychological literature on microaggressions should be understood as part of what Nyla Branscombe and colleagues call the "psychology of the historically disenfranchised" (1999, 135, 146): empirical investigations that focus on the psychology of oppressed social groups rather than, like much of the implicit bias literature, the mental states of those who are privileged.

And it is hypothesized that the aggregate effect of microaggressions—perceived or otherwise—on their targets is significant, and not only because they cause gratuitous pain or discomfort. Perceived discrimination regarding race, gender, and sexual orientation predicts psychological and somatic health outcomes (Mays et al. 2007; Carter 2007; Herek 2009). Racial gaps in health outcomes in the U.S. are not fully explained by differences in socioeconomic status or self-esteem (Gee et al. 2007a, 2007b). Plausibly, microaggressions play a role in explaining these recalcitrant health gaps, and many discussions of microaggressions are motivated by appeal to various outcome gaps (in health, academic or professional achievement, etc.). The detailed mechanism by which microaggressions contribute to such outcome gaps is not known (Okazaki 2009; Torres et al. 2010), but stress seems to be a mediating factor (Harrell and Taliaferro 2003), complicated by in-group identification, which seems to have a protective effect (Crocker and Major 1989; Branscombe et al. 1999). The scientific situation is made more complicated by the multiplicity of experimental protocols (Sullivan 2009): since microaggression incidence is measured in a variety of ways, experimental inference about microaggressions is complicated in ways that are played down in published literature. And in some discussions, "microaggression" may function as a catchall term referring to any manifestations of structural oppression that are relatively difficult to measure independently.

So in the interest of promoting a little more clarity, we distinguish two MICROAGGRESSION concepts. The *explanatory* MICROAGGRESSION concept refers, ex hypothesi, to some factor that explains recalcitrant gaps in desirable outcomes (e.g. good health, professional success) between members of privileged and oppressed social groups, such as those that remain after other factors like wealth, income, and legal discrimination are accounted for. Microaggressions in this sense may turn out to be a variety of diverse factors (they may be "lumpy"; see Feest 2020); we will not know exactly what they look like until we have a more sophisticated causal understanding of recalcitrant outcome gaps. But the term "microaggression" functions in some discourse as a more determinate label for concrete experiences of slights and invalidations. So, let the *hermeneutical* MICROAGGRESSION concept be what is invoked in such contexts.

The hermeneutical MICROAGGRESSION concept is a hermeneutical resource (Fricker 2007) that helps people to make sense of their lived experiences, and the popularity of the MICROAGGRESSION concept outside of the behavioral and social sciences is probably largely due to its hermeneutical role. It is an open empirical question whether the explanatory and hermeneutical MICROAGGRESSION concepts are largely coextensive.

3. Two senses of "subjective." So are microaggressions objective? The answer depends on whether we are talking about explanatory or hermeneutical microaggressions. But clarification is also in order regarding the terms "objective" and "subjective." Philosophers tend to reserve the term "subjective" for propositions whose truth values vary according to a perspective (MacFarlane 2014: a "context of assessment"). For example, a dress may look blue and black to me, and may look white and gold (i.e. not-blue-and-black) to you. There is a perspective-independent fact about what color the dress *is*, but no such fact about how the dress *looks*; it looks different to different people. Let us call such claims *alethically* subjective, and claims that have perspective-independent truth values can be called alethically objective.

By contrast, in common parlance a claim is often said to be "subjective" if reasonable people disagree about its truth value, even if the claim has a perspective-independent truth value. We may call claims that are controversial in this manner *discursively subjective*. The claim that Shakespeare's works were written by William Shakespeare is discursively subjective—some folks believe the plays and poems were written by someone else. But there is a perspective-independent fact of the matter about who wrote Shakespeare's works, so the claim is not alethically subjective.[2] Both alethic and discursive subjectivity are properties of claims rather than concepts or words, but for ease of expression we will talk about "microaggressions" as subjective or objective, meaning that classifying an act or state of affairs as a microaggression is subjective or objective.

Now, obviously claims about microaggressions can be discursively subjective—there is often disagreement about whether a particular act or state of affairs is a microaggression. Nevertheless, it is commonly held that microaggressions are also alethically subjective. Lilienfeld criticizes the microaggression concept on the grounds

[2] Another sense of "objectivity" relevant to science is independence from values or normative commitments, but most microaggressions research is plausibly not objective in this sense since it presupposes a normative theory of justice and structural oppression. However, discussion of the value-free ideal in the social sciences is beyond the scope of our argument.

that microaggressions are thought to be "necessarily in the eye of the beholder" (2017, 143), and Sue claims that "Microaggressions are about experiential reality" (2017, 171). Lilienfeld regards the subjectivity of microaggressions as a source of confusion:

> If Minority Group Member A interprets an ambiguous statement directed toward her […] as patronizing or indirectly hostile, whereas Minority Group Member B interprets it as supportive or helpful, should it be classified as a microaggression? The [microaggressions] literature offers scant guidance in this regard. (Lilienfeld 2017, 143)

Generally speaking, that a claim is discursively subjective does not imply that it is alethically subjective (e.g. the Shakespeare case above is discursively subjective but not alethically subjective). So even if there is reasonable disagreement about whether a particular act or state of affairs is a microaggression, that does not imply that the microaggression is alethically subjective.

Lilienfeld continues:

> it is unclear whether any verbal or nonverbal action that a certain proportion of minority individuals perceive as upsetting or offensive would constitute a microaggression. Nor is it apparent what level of agreement among minority group members would be needed to regard a given act as a microaggression. (Lilienfeld 2017, 143)

Such questions are unmotivated. No serious proponent of the MICROAGGRESSION concept holds that poll results should determine which acts are microaggressions. While "focus groups" and similar methods are sometimes used to determine which kinds of acts should be regarded as microaggressions (e.g. the use of Consensual Qualitative Research methods in Nadal et al. 2015), researchers do not assume that intersubjective agreement among participants is a criterion for being a microaggression. Rather, "focus group" methods are generally employed as techniques for discovering new varieties of microaggression while minimizing the role of researcher biases (see e.g. Nadal et al. 2015, 150–151).

Furthermore, as we argue below, there is no a priori reason to regard microaggressions as alethically subjective. Regarding explanatory microaggressions, it is an open empirical question whether outcome gaps are explained by *perceived* microaggressions or by microaggressions *regardless of how they are perceived by their targets* (i.e. microaggressions ascribed according to an alethically subjective or objective

criterion), or by some other factor. Regarding hermeneutical microaggressions, the concept fails to serve as an adequate hermeneutical resource unless microaggressions are regarded as objective.

4. Explanatory microaggressions: measures and constructs. Lilienfeld observes that most microaggression studies rely on self-report measures, and takes this to be a consequence of the fact that microaggressions are alethically subjective (2017, 151). For example, many studies of microaggressions against African Americans use an instrument called the Daily Life Experiences (DLE) scale (e.g. Scott 2003; Seaton et al. 2009; Torres et al. 2010), developed by Jules Harrell. The instrument consists of 17–20 items describing discriminatory experiences, such as "overhearing or being told an offensive joke" or "being left out of conversations or activities" (from Seaton et al.). Study participants rate how often they have each kind of experience on a scale from "never in the past year" to "once a week or more." Their responses are analyzed (in various ways, depending on the study) to obtain a quantity representing how often participants experience racial microaggressions. The DLE scale is a so-called "self-report" or "subjective" measure, since study participants more or less transparently report information in which experimenters are interested for its own sake (in contrast to behavioral measures or other indirect measures). Self-report measures are common in psychological research on "subjective" constructs like subjective well-being (Alexandrova 2008) or conscious visual experience (Boone 2013), where a "construct" in psychology is a theoretical term whose quantity can be measured (Stone 2019, 1250 n2).

However, the connection between subjective constructs and so-called "subjective measures" is not straightforward. An experimental measure will generally differ from its associated construct in various ways. For example, a Stroop test may be administered as a measure of cognitive depletion (as in e.g. Richeson and Trawalter 2005). But Stroop performance is a temporal measure (a relative delay, measured in milliseconds) whereas cognitive depletion is theoretically something more abstract: it may manifest as a temporal delay or as poorer performance or in various other ways. So here a temporal measure is used to approximate, for the purposes of experimental analysis, the quantity of a more abstract construct of interest (cognitive depletion).

More to the current point, self-report measures may be used to gather information about constructs whose values are alethically objective. Consider, for example, the Perceptual Awareness Scale (PAS), a graded measure of visual awareness (Ramsøy and Overgaard 2004). Study participants are briefly shown an image (often for less than 250

ms) and asked to classify their visual experience as "Clear Image," "Almost Clear Image," Weak Glimpse," or "Not Seen." One may think that this is a subjective measure for a subjective construct, since visual experience is often said to be "subjective," but alethic subjectivity is a property of claims so we must be precise about what claim is at issue. Visual experience is subjective in that the content of two visual experiences may differ for various judges (or one judge at different times) although those experiences are of the same object in the same conditions. We often characterize the contents of such experiences using clauses with "seem" or "look" as the main verb, and such clauses are alethically subjective. A dress may look blue to Ali and at the same time look white (i.e. not-blue) to Leah; the truth value of an utterance like "This dress looks blue" may vary depending on the judge. But the PAS does not measure what the content of a visual experience is; the PAS measures whether a visual experience of a stimulus occurred for a particular observer, and how clear that experience was. This is an alethically objective state of affairs. The truth value of "Ali had a clear visual experience of the stimulus" does not vary according to who evaluates it. If Ali and Leah disagree about the truth of such a sentence, then one of them must be wrong (and it's probably not Ali).

Similarly, instruments like the DLE scale, which purport to reveal rates of microaggression incidence in a participant's life through self-report, may be fallible measures of an alethically objective quantity. We say "may" because much extant microaggression research does not distinguish clearly between alethically objective and subjective interpretations of microaggression incidence. Instruments like the DLE scale may be used either to measure the frequency of a participant's exposure to demeaning incidents (an alethically objective quantity), or to measure the participant's perception of how often she experiences demeaning incidents (an alethically subjective quantity). Microaggressions in the explanatory sense are some factor that explains recalcitrant gaps in desirable outcomes between members of privileged and oppressed social groups, such as those that remain after other factors like wealth, income, and legal discrimination are accounted for. It is an open question whether this factor is (1) mere exposure to demeaning incidents, regardless of how they are perceived by their targets, or (2) the perception of one's experiences as demeaning, or (3) something else. That is, it is an open empirical question whether explanatory microaggressions are alethically objective or subjective. Further empirical study is needed to assess the relative merits of these hypotheses.

As a matter of verbal hygiene, it seems reasonable to us to treat explanatory microaggressions as alethically objective, and then to examine whether outcome gaps are caused by exposure to microaggressions per se or by the perception of events as

microaggressions. By analogy, the standard for whether an act is a sexual assault is not whether the survivor characterizes the act as "sexual assault" or even as harmful. But the matter of which way to speak can only be settled by the community of speakers (in this case, the community of social and behavioral scientists), not by fiat, and it seems to us that the matter has not yet been settled.

We wish to be clear that microaggression research employs a variety of methods that vary in quality and purpose; the DLE scale is only one instrument among many. Our main objective here is not to conduct a methodological review (for which see e.g. Okazaki 2009; Lau and Williams 2010; Wong et al. 2014), but to argue against a tempting error. It is simplistic to identify a construct with its measure, and it is an error to freely attribute the properties of a measure to its associated construct. So while it is true that microaggression frequency is often measured using participant self-reports, we should not infer from this fact that microaggression incidence is alethically subjective. Existing measurement practices do not settle the question of whether microaggressions are "in the eye of the beholder."

5. Hermeneutical microaggressions: phenomenological considerations. Whereas it is an open question whether explanatory microaggressions are alethically subjective, there is compelling reason to regard hermeneutical microaggressions as alethically objective. Our argument depends on the commonly reported phenomenology of microaggression targets. Members of oppressed groups often report experiencing confusion and uncertainty about whether an act directed toward them is a subtle expression of prejudice, or whether it is no different than an act that would have been directed toward a privileged person. This feature of microaggressions is sometimes called "attributional ambiguity" (Crocker and Major 1989). For example, a woman might be addressed at work by her first name (e.g. "Stephanie") rather than by her title and surname (say, "Dr. Appiah"). In a context where either form of address is acceptable, and where the base rates are not known (i.e. it is not known how often people in general, or people of various genders, are addressed by their first names vs. by their titles and surnames), it can be difficult to determine whether the address expresses a slight.

Here is an argument that we should consider hermeneutical microaggressions to be alethically objective. Supposing the contrary, that microaggressions are alethically subjective, there are two possibilities. First, perhaps, as in many matters of taste, it is appropriate to allow everyone their own perspective. So whoever feels the act of addressing the woman by her first name was a gendered slight regards it as a

microaggression, and whoever feels the form of address was not influenced by gender does not regard it as a microaggression. If microaggressions are alethically subjective, as we are currently supposing, then there is no perspective-independent fact of the matter about whether this incident is a microaggression (as in predications of "is tasty" or "looks blue-ish to me"). A second possibility is that people have their own perspectives but the target's perspective is decisive: the act is a microaggression if and only if the addressed woman feels slighted. In both of these possibilities, it makes no sense for the woman to wonder whether the act was really an expression of prejudice, i.e. whether it was really a microaggression. On the first option, there is no fact of the matter about whether the act was a microaggression. On the second option, the matter is decided by the woman's own perspective, so her judgment settles the question.

However, people who experience relatively subtle microaggressions often report wondering precisely about this. Indeed, it is often claimed (e.g. by Sue et al. 2007; Bartky 1975; Du Bois 1903 and others) that much of the harm of microaggressions is caused precisely by anxiety and paranoia regarding one's inability to quickly and accurately assess whether an act was indeed a microaggression. Only the objectivist view of microaggressions accounts for this phenomenology. If we seek hermeneutical justice, we have reason to adopt concepts that make sense of rather than obscure common experiences for members of oppressed social groups (Fricker 2007). So we should regard microaggressions as objective, in that there are perspective-independent facts about whether particular acts or states of affairs are microaggressions in the hermeneutical sense.

6. Conclusion. We argued, against the common view, that microaggressions should not be regarded as alethically subjective. For microaggressions in the hermeneutical sense—considered as a category of items that help members of oppressed social groups to make sense of their lived experience—we argue that only an objectivist view rationalizes the distress commonly experienced due to attributional ambiguity. For microggressions in the explanatory sense—considered as the causes of recalcitrant outcome gaps—we acknowledge that it is an open question whether they are best regarded as alethically objective or subjective. But we argued against a tempting view, expressed by Lilienfeld and others, that self-report measures are especially suited for measuring the value of theoretical constructs that are alethically subjective. People will continue to question whether particular acts or states of affairs count as

microaggressions, and we contend that those questions have objectively accurate responses.

REFERENCES

Alexandrova, Anna. 2008. "First-Person Reports and the Measurement of Happiness." *Philosophical Psychology* 21: 571–583.

Barthelemy, Ramón S., Melinda McCormick, and Charles Henderson. 2016. "Gender Discrimination in Physics and Astronomy: Graduate Student Experiences of Sexism and Gender Microaggressions." *Physical Review Physics Education Research* 12 (020119): 1–14.

Bartky, Sandra Lee. 1975. "Toward a Phenomenology of Feminist Consciousness." *Social Theory and Practice* 3: 425–439.

Boone, Worth. 2013. "Operationalizing Consciousness: Subjective Report and Task Performance." *Philosophy of Science* 80: 1031–1041.

Branscombe, Nyla R., Michael T. Schmitt, and Richard D. Harvey. 1999. "Perceiving Pervasive Discrimination among African Americans: Implications for Group Identification and Well-Being." *Journal of Personality and Social Psychology* 77: 135–149.

Capodilupo, Christina M., Kevin L. Nadal, Lindsay Corman, Sahran Hamit, Oliver B. Lyons, and Alexa Weinberg. 2010. "The Manifestation of Gender Microaggressions." In Sue 2010b, 193–216.

Carter, Robert T. 2007. "Racism and Psychological and Emotional Injury: Recognizing and Assessing Race-Based Traumatic Stress." *The Counseling Psychologist* 35: 13–105.

Crenshaw, Kimberle. 1989. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." *University of Chicago Legal Forum* 1989: 139–167.

Crocker, Jennifer, and Brenda Major. 1989. "Social Stigma and Self-Esteem: The Self-Protective Properties of Stigma." *Psychological Review* 96: 608–630.

Du Bois, W.E.B. 1903. *The Souls of Black Folk*. Chicago: McClurg.

Feest, Uljana. 2020. "Construct Validity in Psychological Tests – the Case of Implicit Social Cognition." *European Journal for Philosophy of Science* 10. https://doi.org/10.1007/s13194-019-0270-8.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Gee, Gilbert C., Michael Spencer, Juan Chen, Tiffany Yip, and David T. Takeuchi. 2007a. "The Association between Self-Reported Racial Discrimination and 12-Month Dsm-Iv Mental Disorders among Asian Americans Nationwide." *Social Science and Medicine* 64: 1984–1996.

———. 2007b. "A Nationwide Study of Discrimination and Chronic Health Conditions among Asian Americans." *American Journal of Public Health* 97: 1275–1282.

Gonzalez, Lauren, Kristin C. Davidoff, Kevin L. Nadal, and Philip T. Yanos. 2015. "Microaggressions Experienced by Persons with Mental Illness: An Exploratory Study." *Psychiatric Rehabilitation Journal* 38: 234–241.

Harrell, Jules P., and James Taliaferro. 2003. "Physiological Responses to Racism and Discrimination: An Assessment of the Evidence." *American Journal of Public Health* 93: 243–248.

Herek, Gregory M. 2009. "Hate Crimes and Stigma-Related Experiences among Sexual Minority Adults in the United States: Prevalence Estimates from a National Probability Sample." *Journal of Interpersonal Violence* 24: 54–74.

Johnston, Marc P., and Kevin L. Nadal. 2010. "Multiracial Microaggressions: Exposing Monoracism in Everyday Life and Clinical Practice." In Sue 2010b, 123–144.

Keller, Richard M., and Corinne E. Galgay. 2010. "Microaggressive Experiences of People with Disabilities." In Sue 2010b, 241–267.

Lau, Michael Y., and Chantea D. Williams. 2010. "Microaggression Research: Methodological Review and Recommendations." In Sue 2010b, 313–336.

Liao, Shen-yi, and Bryce Huebner. 2020. "Oppressive Things." *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12701.

Lilienfeld, Scott O. 2017. "Microaggressions: Strong Claims, Inadequate Evidece." *Perspectives on Psychological Science* 12: 128–169.

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

Mays, Vickie M., Susan D. Cochran, and Namdi W. Barnes. 2007. "Race, Race-Based Discrimination, and Health Outcomes among African Americans." *Annual Review of Psychology* 58: 201–225.

Nadal, Kevin L., Kristin C. Davidoff, Lindsey S. Davis, Yinglee Wong, David Marshall, and Victoria McKenzie. 2015. "A Qualitative Approach to Intersectional Microaggressions: Understanding Influences of Race, Ethnicity, Gender, Sexuality, and Religion." *Qualitative Psychology* 2:
147–163.

Nadal, Kevin L., Marie-Anne Issa, Katie E. Griffin, Sahran Hamit, and Oliver B. Lyons. 2010. "Religious Microaggressions in the United States: Mental Health Implications for Religious Minority Groups." In Sue 2010b, 287–310.

Nadal, Kevin L., David P. Rivera, and Melissa J.H. Corpus. 2010. "Sexual Orientation and Transgender Microaggressions: Implications for Mental Health and Counseling." In Sue 2010b, 217–240.

Okazaki, Sumie. 2009. "Impact of Racism on Ethnic Minority Mental Health." *Perspectives on Psychological Science* 4: 103–107.

Olkin, Rhoda, H'Sien Hayward, Melody Schaff Abbene, and Goldie VanHeel. 2019. "The Experiences of Microaggressions against Women with Visible and Invisible Disabilities." *Journal of Social Issues* 75: 757–785.

Pierce, Chester. 1970. "Offensive Mechanisms." In *The Black Seventies*, edited by Floyd B. Barbour, 265–282. Boston: Porter Sargeant.

Ramsøy, Thomas Zoëga, and Morten Overgaard. 2004. "Introspection and Subliminal Perception." *Phenomenology and the cogntiive sciences* 3: 1–23.

Richeson, Jennifer A., and Sophie Trawalter. 2005. "Why Do Interracial Interactions Impair Executive Function? A Resource Depletion Account." *Journal of Personality and Social Psychology* 88: 934–947.

Rini, Regina. 2018. "How to Take Offense: Responding to Microaggression." *Journal of the American Philosophical Association* 4: 332–351.

Scott, Lionel D. 2003. "The Relation of Racial Identity and Racial Socialization to Coping with Discrimination among African American Adolescents." *Journal of Black Studies* 33: 520–538.

Seaton, Eleanor K., Tiffany Yip, and Robert M. Sellers. 2009. "A Longitudinal Examination of Racial Identity and Racial Discrimination among African American Adolescents." *Child Development* 80: 406–417.

Smith, Laura, and Rebecca M. Redington. 2010. "Class Dismissed: Making the Case for the Study of Classist Microaggressions." In *Microaggressions and Marginality: Manifestation, Dynamics, and Impact*, edited by Derald Wing Sue, 269–285. Hoboken, New Jersey: Wiley.

Stone, Caroline. 2019. "A Defense and Definition of Construct Validity in Psychology." *Philosophy of Science* 86: 1250–1261.

Sue, Derald Wing. 2010a. Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation. Hoboken, New Jersey: Wiley.

———, ed. 2010b. Microaggressions and Marginality: Manifestation, Dynamics, and Impact. Hoboken, New Jersey: Wiley.

———. 2010c. "Microaggressions, Marginality, and Oppression: An Introduction." In Sue 2010b,
3–22.

———. 2017. "Microaggressions and "Evidence": Empirical or Experiential Reality?" *Perspectives on Psychological Science* 12: 170–172.

Sue, Derald Wing, Christina M. Capodilupo, Gina C. Tonino, Jennifer M. Bucceri, Aisha M.B. Holder, Kevin L. Nadal, and Marta Esquilin. 2007. "Racial Microaggressions in Everyday Life: Implications for Clinical Practice." *American Psychologist* 62: 271–286.

Sullivan, Jacqueline A. 2009. "On the Multiplicity of Experimental Protocols." *Synthese* 167: 511–539.

Torres, Lucas, Mark W. Driscoll, and Anthony L. Burrow. 2010. "Racial Microaggressions and Psychological Functioning among Highly Achieving African-Americans: A Mixed-Methods Approach." *Journal of Social and Clinical Psychology* 29: 1074–1099.

Washington, Natalia, and Daniel Kelly. 2016. "Who's Responsible for This? Moral Responsibility, Externalism, and Knowledge About Implicit Bias." In *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, edited by Jennifer Saul and Michael Brownstein. Oxford: Oxford University Press.

Wong, Gloria, Annie O. Derthick, E.J.R. David, Anne Saw, and Sumie Okazaki. 2014. "The *What*, the *Why*, and the *How*: A Review of Racial Microaggressions Research in Psychology." *Race and social problems* 6: 181–200.

Classificatory norms in scientific practice: the unobjective but rational *chemical element*.

[version accepted for presentation at PSA 2020; please don't cite or quote without authors' permission]

Matthew J. Barker, Concordia University

matthew.barker@concordia.ca

and

Matthew H. Slater, Bucknell University

mhs016@bucknell.edu

Abstract

It is often presumed that empirical considerations provide epistemic objectivity for claims about the boundaries and classification of scientific categories. This has seemed especially plausible in chemistry. Focusing on the category *chemical element*, we describe two 20th century developments that undermine epistemic objectivism about it. But our second thesis is that, in practice, this shortfall is bridged by relying on a little-recognized species of pragmatic norm: classificatory norms. We contend this precludes the objectivity, yet ironically affords the rationality, of related category and classification claims.

1. Introduction.

How in practice do scientists determine where to draw category boundaries?

Compared to many of the questions that philosophers ask about scientific categories and classification, that one gets little attention. It is more epistemic (or methodological) than, say, widely-discussed metaphysical questions about scientific categories. We are interested in it partly because we think addressing it with a focus on scientific practice bodes ill for a range of objectivist metaphysical positions. But the epistemic question is interesting for other reasons too, and here we'll restrict our focus to it, and the challenges it presents to a typically unexamined *epistemic objectivism* about scientific categories and classification.

We focus on practice in chemistry, a domain often regarded as a bastion of classificatory objectivity[1]—in particular, on how views about the category *chemical element* were (and were not) defended during two 20th century episodes. This will allow us to argue for an epistemic anti-objectivism that is surprising partly because it still allows for proposals about category

[1] Famous examples outside philosophy of chemistry include Putnam and Kripke (Putnam 1975; Kripke 1980); within present philosophy of chemistry, Scerri is a well-known objectivist about *chemical element*. Note such popular objectivisms are compatible with conventionalisms about the Periodic Table of Elements, which have become wide-spread (Scerri 2007, 277–78).

boundaries to be more or less *rational* in virtue of the operation of classificatory norms that we'll uncover and describe.

More specifically, we'll argue for two theses. First:

Short-Fall: sometimes, empirical considerations alone fall short of stance-independently justifying theories about which conditions are the constitutive ones for a scientific category.[2]

Our second thesis is related:

Bridging-By-Norms: in some cases of classificatory short-fall, scientists bridge the epistemic gap by relying on classificatory norms.[3]

---

[2] Philosophers of science have long discussed the extents to which empirical considerations leave theory choice undetermined. But there has been scant attention in this literature to theories about category constitution in particular, which differ in various ways from the theories usually investigated.

[3] A typical classificatory norm is a pattern of classificatory *behavior or belief* that stems from some people *preferring*, implicitly or explicitly, to behave or believe in such-and-such a way under certain conditions. (See Bicchieri (2017) for such a view of norms.) Consequently, classificatory norms differ from extra-empirical virtues, which are more like favored *properties of theories*. Unlike extra-empirical virtues, specifically *classificatory* norms have received scant notice or investigation.

In addition to arguing for those two theses, we will briefly remark on the prospect for reasonable appeals to classificatory norms that allow for some proposed category boundaries (and some related classificatory claims) to be more rational than others. But prior to arguing for our theses, let us clarify them.

2. Clarifications of Short-Fall and Bridging-By-Norms Theses.

By 'empirical considerations' we include appeals to observational data, and to theories that are widely deemed highly confirmed.

Regarding a category's constitutive conditions, we mean those conditions in virtue of which (in usual circumstances) a thing satisfying them belongs to that category.[4] Some metaphysical objectivisms imply that for many scientific categories, which conditions are constitutive of belonging is an objective matter. The epistemic objectivism more relevant here is about the supposed justifications of theories about constitutive conditions of categories. It says that in some and perhaps many cases, the considerations advanced in support of the theory suffice to objectively justify the truth of its proposals about which conditions are constitutive. We will presume such considerations objectively justify a theory about

[4] Using our terminology, the category *chemical element* is supposed to be a piece of the world that science attempts to track with the concept CHEMICAL ELEMENT.

constitutive conditions *only if* the favoring they provide the theory is *independent of our mere mental stances* towards those considerations.

One can fail to meet that necessary condition on objective justification in obvious ways, such as via wishful thinking, where one *wants* certain considerations to favor the truth of a particular theory about category constitution, while having little or no evidence or reason to complement that desire. In our chemical cases, wishful thinking isn't an issue. Something much less obvious is going on. Chemists are, perhaps very rationally, relying on widespread implicit norms in order to support their theories about the constitutive conditions of *chemical element*. To the extent that relying on such norms involves relying on various mental stances of peers within a scientific community (see Bicchieri 2017), it precludes the objectivity of justification in question.

3. Lumpers vs. Splitters, ~1910–1920s.

The first of the two 20th century episodes we investigate played out in publications and meetings from about 1910 into the 1920s, in the wake of emerging details about atomic structure. There was a dispute about, roughly, whether and in what way the discovery of isotopy should revise Mendeleev's 19th century view that each *place* in the Periodic Table represents exactly one distinct chemical element. By 1910 that view was widespread, despite questions remaining about the exact *sequences* of elements within the Table (Scerri 2007). We

can understand those who then took isotopy to then challenge Mendeleev's view as splitters, and those who defended it as lumpers.

To see this, consider some context and details. Frederick Soddy is often credited with discovering isotopy.[5] He first proposed the idea in 1911 and introduced the term 'isotope' to chemistry in 1913.[6] He did not describe isotopy in terms of protons or neutrons, because neither had been discovered yet. He based his proposal largely on investigations of decay chains that indicated more than 30 different species of element, called "radioelements", over a stretch of the Periodic Table where just 11 elements were so far acknowledged (Choppin et al. 2013). Each of these radioelements was then said to be an isotope, with "mesothorium" and "thorium X" as examples (Scerri 2007, 177). Nowadays we regard each of these as isotopes of radium—as mere variants of that element. But when Soddy proposed the existence of isotopes, some researchers, especially the radiochemist and discoverer of protactinium Kazimierz Fajans, urged that each isotope was its own chemical element (Scerri 2000). Researchers like Fajans were thus splitters in the sense that they saw some places in the Periodic Table as subsuming or splitting into multiple elements rather than representing just one element each. Those who resisted this while nonetheless granting the existence of isotopes, e.g., those who

[5] Others are recognized as anticipating aspects of it, including William Crookes as early as 1886 (Scerri 2007, 176).

[6] As Scerri notes (2007, 312), Soddy got the term from Margaret Todd.

grouped isotopes such as “thorium X” and “mesothorium” together as mere varieties of one element, can be understood as lumpers.

An implication is that splitters and lumpers were operating with incompatible theories about the constitutive conditions of the category *chemical element*. It is difficult to pin down these theories because they were in flux and usually implicit rather than explicit throughout the period of opposition. But we can get far enough to see how the theories support our Short-Fall thesis.

What made for the differences between places in the Periodic Table? More than 30 years earlier, Mendeleev had thought the answer was a mix of differing atomic weights and chemical properties. But by 1910, physicists were using electron scattering experiments to investigate the structure of chemical constituents. Subsequently, as Scerri helpfully recounts (2007), over the course of the next 13 years several researchers—including Rutherford, Barkla, van den Broek, Moseley, and Chadwick—used and developed this and related work to motivate a shift, from understanding chemical element identity in terms of atomic weights and chemical properties, to understanding it in terms of *atomic number* equated with (an early notion of) positive nuclear charge. In making that idea explicit in its 1923 definition of ‘chemical element’ (Aston et al. 1923), the IUPAC was stating a view that had been implicitly held by many chemists since the work of van den Broek and Moseley 10 years earlier.

Summarizing these developments, we can say that between 1913[7] and 1923 lumpers widely recognized the following theory of the constitution of the *chemical element* category:

Positive Charge Theory:

Any thing is a chemical element if and only if:

(a) it is a category, a species, of atom,[8] and

(b) all atoms of this species have the same atomic number, which = nuclear positive charge, and

(c) only atoms of this species have that atomic number.

Splitters such as Fajans rejected this when urging that isotopes of the same atomic number are each distinct elements in their own right, which effectively denied part (c) of what we've termed the Positive Charge Theory.

[7] This year for the theory rather than 1910 because it wasn't until 1913 that van den Broek had finished disconnecting the identity of atomic number from atomic weight in favor of atomic charge.

[8] Although part (a) of the definition now seems unremarkable, there are complications (Scerri 2000), which were influentially discussed by Paneth (e.g., Paneth 1962a; 1962b). Although splitters sometimes appeal to these complications, their position didn't require them and we'll set them aside here.

In light of this opposition, recall our Short-Fall thesis. It says that sometimes empirical considerations alone fall short of stance-independently justifying theories about which conditions are the constitutive ones for a scientific category. We'll now argue this was the case in the opposition between lumpers who supported the Positive Charge Theory and splitters who rejected it.

Empirical results were certainly relevant. A rapid succession of detected differences between radioelement isotopes in decay studies, for example, fueled dispute (Scerri 2007). However, no such empirical findings, on their own, stance-independently justified either accepting or rejecting (c). This may sound odd, given that by the 1930s virtually all chemists were lumpers and today you would earn incredulous stares if you proposed that isotopes of the same atomic number are distinct chemical elements. But this paradigm example of classificatory consensus owes in part, we submit, to widespread implicit agreement on classificatory norms—not just to impressive empirical findings.

To appreciate this, consider how splitters dug in their heels even when lumpers generated impressive empirical results that seemed to favour lumping. One set of such results were negative—the inability, despite repeated attempts, to *chemically* distinguish the isotopes that were being discovered (Scerri 2007, 177). Another set were positive—showing extensive chemical *similarities* between isotopes of shared atomic number. As just one example, Fritz Paneth and György von Hevesy reported on electrochemical experiments in 1914 that "observed voltage was found to be constant, regardless of the proportion of the two isotopes [of bismuth] present in the sample" (Scerri 2000, 63). A main way that splitters objected was

by contesting the results. Fajans, for instance, disputed the bismuth results and insisted against Paneth and von Hevesy that the compared isotopes were distinct elements (Scerri 2000, 65).

This has the surface appearance of making this moment in the dispute turn solely on an empirical matter. But one very probable reason why Fajans contested the empirical results was that he roughly shared with his opposition a norm that leant classificatory relevance to the results—something like:

> *Elemental Relevance Norm*: If you are determining whether different isotopes are instances of the same chemical element, and the empirically detected differences between them seem small or unimportant in comparison to the empirically detected similarities between them, then judge that the differences lack elemental relevance and the isotopes are instances of the same chemical element.

It would have been odd for Fajans to contest the empirical results were he not presuming something like the view captured in that norm. Why worry (as he did) about reported *similarities* if you're not basing your classification claims on some judgment about the *relevance of* reported similarities vs. differences?

Something like the Elemental Relevance Norm also helps clarify the importance that a distinction between chemical and physical properties was eventually deemed to have in these debates. Even Fajans eventually bowed somewhat to this distinction (Scerri 2000, 64). Researchers also began appreciating more classificatory *relevance* in the distinction. They can be understood as doing this via the Elemental Relevance Norm, where the physicalness, so-to-speak, of differences is presumed to give a reason for counting those differences as small or

unimportant in comparison to the many similarities deemed chemical. Seen in this light, the role of that norm has probably increased significantly over time, as we've retained what is effectively a lumper's view of *chemical element* despite discovering many further physical differences between isotopes of atoms that share their atomic number. Norms like the Elemental Relevance Norm allow people to acknowledge that *physical* differences between uranium isotopes, for instance, are relevant—even dramatically important—to a great many things (including the energy industry and warfare), without conceding that the very specific issue of the constitutive conditions of a *chemical* category is one of those things.

The Elemental Relevance Norm also appears alive and well today, as claims that the recently increasing number of known chemical differences between isotopes are not yet numerous or important enough to challenge lumpers (e.g., Scerri 2007, 279, 327) seem most charitably interpreted as implicitly involving reliance on that norm.

Of course, none of this is to criticize any particular stance on or use of the norm, nor any associated relevance assumptions or claims. Indeed, we reckon that reliance on the norm was and continues to be quite rational, an issue we return to below. But recognizing the rationality or wisdom of a norm's role is to already grant its operation alongside empirical data, which is our point here. In effect, by zeroing in on one dispute about the category *chemical element* we have supported our Short-Fall thesis by elaborating and supporting our Bridging-By-Norms thesis.

4. Protons vs. Electrons, ~1930s–Present.

The second 20th century episode we discuss is not so much an explicit dispute as it is an implicit opposition between two theories about the constitutive conditions of *chemical element*, only one of which is explicitly wide-spread. It is an opposition that most experts do not even acknowledge, even though their explanatory and classificatory practices generate it. It starts from a curious relationship between two underlying views about elements that are perhaps as close to unanimous as two views can get.

The first of these underlying views, and the acceptance of it, are indicated in the fact that nearly all experts today hold to a descendant version of the 1923 IUPAC definition of 'chemical element'. In light of what was learned about proton counts later than 1923, this descendant version defines 'chemical element' explicitly in terms of those counts: a chemical element is, in the sense in question, "a species of atoms", where each of these species is made up of "all atoms with the same number of protons in the atomic nucleus" (IUPAC 2019).

Correspondingly, the received theory about the nature of the *chemical element* category is that atomic proton count is *the* central condition—it is what ontologically makes an atom the kind of element it is. In table 1 we've summarized this more precisely as the *Proton Count Theory*, using single asterisk marks to indicate parts of the theory that contain revisions to the Positive Charge Theory that is associated with the older (1923) IUPAC definition of 'chemical element'. The revisions simply involve referring to proton count rather than nuclear positive charge.

Table 1

Three Different Theories about the Constitutive Conditions of the Category *Chemical Element*

Proton Count Theory	Electron Configuration Theory	Proton & Electron Theory
Any thing is a chemical element if and only if: (a) it is a category, a species, of atom, and (b*) all atoms of this species have the same atomic number, which = proton count, and (c*) only atoms of this species have *that* atomic number.	Any thing is a chemical element if and only if: (a) it is a category, a species, of atom, and (b**) all atoms of this species have the same ground state electron configuration, which = … (c**) only atoms of this species have *that* ground state electron configuration.	Any thing is a chemical element if and only if: (a) it is a category, a species, of atom, and (b***) all atoms of this species have the same atomic number, which = proton count, and the same ground state electron configuration, which = … (c***) only atoms of this species have *that* atomic number, and *that* ground state electron configuration.

NOTE.— For simplicity we have not fully elaborated conditions (b**) and (b***), which would involve referencing, e.g., the first, second, third, and fourth quantum numbers, the Pauli exclusion principle, the Aufbau principle, and the Hund principle (Scerri 2007, 233ff.).

The second unanimous (or very nearly so) underlying view, existing alongside the Proton Count Theory, is about explanation. As Scerri succinctly puts it, "it is the electron that is mainly responsible for the chemical properties of the elements" (Scerri 2007, 160). In other words, the main thing that explains—causally or otherwise—most properties and behaviors of respective chemical elements is their respective electron configurations. This isn't to say that proton count has *zero* role to play in such explanations. Number of protons helps determine and influence atomic forces and structure, and interacts with electron configuration and behavior, and these things also *help* explain features and behaviors of chemical elements, e.g., why atoms of sodium together react as they do with atoms of chlorine. And of course, other variables aside from just proton count and electron configuration are also parts of any *complete* explanations of elemental features and behaviors. But the resounding view is that electron configurations do most of the explaining, with other variables often being negligible—hence the preceding quote from Scerri.

So when it comes to *explanations* about elements in reactions, electrons are deemed central. But when it comes to element *identity*, and so the constitutive conditions of the *chemical element* category, the consensus is that proton count is central and there is no reference to electrons.

This is curious because it seems to conflict with a norm that prioritizes explanatorily-central conditions when theorizing about a category's constitutive conditions—a norm that operates in many other areas of science. It may be that this norm is popular with respect to

categories of a certain general sort, so we should first clarify this and how the *chemical element* category seems to be of that sort.

Many categories consist in *patterns of linked variables*. Disease categories are a vivid example, with one disease often being distinguished from others by how it consists in recurring (across cases) linkages between two types of variables: characteristic symptoms or effects, on one hand, and their causes, on the other. Recent work clarifies that other biological categories are like this too (e.g., [Suppressed-for-review]). Although these examples involve cause and effect variables, key variables may be of other sorts, that is, with determination relations other than causation between them.

The category *chemical element* seems a paradigm example of a linked variable category. It appears to consist in a set of distinct patterns of linkage—some associated with one element, others with other elements—between particular sets of chemical properties or behaviors, and the conditions in each case that are mainly responsible for bringing about those properties and behaviors. What makes these patterns alike are the sorts of variables involved. Whether we're talking about the element *chlorine*, or *gold*, or *hafnium*, etc., there are certain types of links between the chemical properties of those elements, on one hand, and the conditions responsible for those properties, on the other. Chlorine displays certain properties of reactivity due to conditions involving electron configuration. Gold displays different properties of reactivity, but similarly due to conditions involving (different) electron configuration.

Now here is the associated implicit norm that seems widespread:

> *Main Explanatory Variables Norm*: If you are determining which conditions are constitutive of a category that consists in patterns of linked variables, then prioritize those conditions that are the main explanatory factors for the other variables within the patterns in question.

Were experts spelling out conformity with this norm in the *chemical element* case, they would note what is distinctive about the elements that exemplify the category. Each exhibits patterns of chemical properties and behaviors, explained by a combination of variables. What are the main explanatory variables? According to them: electron configurations. They then would, presumably, propose electron configurations to be constitutive of the category *chemical element*—either fully constitutive or partly constitutive, as represented in the theories stated in the middle and right-hand columns of table 1—by referring to those configurations in their definition of 'chemical element'. But as we have seen, they don't do that.

Why does it seem experts in chemistry don't follow the *Main Explanatory Variables Norm* that is common in many analogous cases in science? Why not connect the issues of explanation and identity? Two different answers come to mind.

One is that perhaps appearances are misleading here—that, actually, chemists *are* following this norm. Perhaps we have expressed the norm in too coarse-grained a way to see this. A more fine-grained version could distinguish between *proximate* and *distal* explanatory variables, allowing appeal to more distal variables to abide the norm. Some authors argue, for instance, that while electron configurations or structure are the main *proximate* variables that explain elemental properties and behaviors, electronic structure is in turn determined or

explained by proton count (nuclear charge) (Hendry 2012, 266). Proton counts may then be the main *distal* variables that explain elemental properties and behaviors, bringing the consensus view about element identity into one kind of harmony with the consensus view on element explanations.[9]

The second type of answer takes appearances at face value, conceding that chemists aren't following the *Main Explanatory Variables Norm*. But if they aren't, it is very probably because they are following *other* norms given priority over that one. One likely other norm in this case would trade on a seemingly perfect correlation between atomic number and ground state electron configuration, along with a penchant for simplicity. This norm grants that electron configurations are the main explanatory variables, but emphasizes that each ground state configuration always corresponds with exactly one atomic number. And appealing to these atomic numbers is simpler than appealing to ground state electron configurations, in that reference to proton count is less complicated and more concise than reference to the quantum numbers and associated quantum mechanical principles (see table 1) that give the ground state electron configurations. This is to recognize three rather than just two linked variables: first, the chemical properties and behaviors, second the electron configurations that explain those

[9] For discussions of chemistry-physics relationships that may provide other grounds for arguing that chemists are abiding the *Main Explanatory Variables Norm* after all, see Scerri (2007).

properties and behaviors, and third the simpler atomic numbers that are explanatorily negligible but which correlate perfectly with the explanatory electron configurations. The norm in question would then imply that when determining constitutive conditions under such circumstances, we should prioritize the third variables, the simpler ones—the atomic numbers in this case—rather than the main explanatory variables with which they correlate:

> *Perfect Correlation Simplicity Norm*: If you are determining which conditions are constitutive of a category that consists in patterns of linked variables, and in addition to one type of variable that is mainly explained by another there is also a third type of variable that is simpler than the explanatory variables but correlates perfectly with them, then prioritize those simpler correlated variables as constitutive of the category in question.[10]

This norm is tempting because it buys helpful simplicity at little cost. Indeed, following it in practice would seem to come with *zero* risk of recognizing different element boundaries than someone who instead recognizes ground state electron configurations as constitutive. Put

[10] Note that this norm is recommending that the simpler variables that correlate with the explanatory ones (but which are not themselves explanatory) be prioritized when specifying constitutive conditions. Some alternative strategies that view the simpler variables as themselves explanatory (or close enough proxies for what is explanatory) would signal operation of the *Main Explanatory Variables Norm*.

differently, the three theories in table 1 certainly differ, but are, so far as we know, coextensive in application.

There may also be other norms that tacitly trump the *Main Explanatory Variables Norm*. It has been widely noted that through the years in which a consensus built around defining 'chemical element' in terms of atomic number, the element concept in chemistry was being used to capture units that survive chemical change—what survives of sodium and chlorine, respectively, for example, when each seems to give way as they combine to form a salt (Scerri 2000). This may seem to privilege atomic numbers over electron configurations because the former survive such reactions while the latter change. Expressing this sort of privilege, for instance, Hendry writes that "whatever earns something membership of the extension 'krypton' must be a property that can survive chemical change, and therefore the gain and loss of electrons" (Hendry 2012, 266). Perhaps this indicates:

> *Unchanging Constituents Norm*: If you are determining which conditions are constitutive of a category, then prioritize those that remain unchanged in persisting category members, over those that change in category members.

That is probably too simplistic as stated though. Only somewhat sloppy adherence to it would in fact privilege the Proton Count Theory over the other two theories in table 1 because an atom's *ground state* electron configuration—the structure its electrons *would* take were it in a neutral ground state—is a *disposition*, sometimes retained when the ground state happens not to obtain, e.g., during chemical reactions. So if experts really do tacitly rely on the *Unchanging Constituents Norm* in a way that trumps the *Main Explanatory Variables Norm*,

then it would probably be a more sophisticated version that implies a preference for unchanging *manifest* properties rather than unchanging dispositional ones.

There are surely other plausible candidates for norms that experts have leaned on if they indeed have opted against the *Main Explanatory Variables Norm*. The overarching point is that either way, norms are involved. If appearances are misleading and at least some chemists have kept elemental explanation and elemental identity connected, it seems they have deployed the *Main Explanatory Variables Norm*; if for some or all chemists the appearance of disconnecting these things is instead accurate, then norms seem to help support that disconnection. Either of those paths to selecting the widely accepted Proton Count Theory over the other theories in table 1 leads through norms in addition to empirical considerations.

5. Conclusion.

Our intention in this short paper is to show how classificatory shortfall (our first thesis) with bridging by norms (our second thesis) occurs even for a chemical category alleged to enjoy a great deal of classificatory objectivity. Uncovering such shortfall bridging in two 20th century episodes in chemistry challenges an epistemic strand to that alleged objectivity, given what the norms in question and epistemic objectivity were clarified in section 2 to involve. Admittedly, this does not constitute an argument that classificatory shortfall with bridging by norms is inevitable (for a more general case, see [suppressed-for-review]). But it may surprise many that it happens at all.

Beyond our main goal of arguing that this happens in surprising contexts, we also briefly noted that the rationality of involved classificatory decisions can survive the loss of objectivity we've documented. To support that as a thesis additional to the two we've argued for here would require another paper. But the prospects should now seem favorable: given the extent to which classificatory norms are shared, can pragmatically aid attainment of goals in a research community, and are continuous with the theories in which they are embedded, the classificatory decisions they guide can be properly seen as rational in a robust sense despite lacking epistemic objectivity. Such rationality may come to seem especially important if the problems we've posed for an epistemic objectivism about categories are found to also cast doubt on more metaphysical objectivisms about them.

References

Aston, F. W., Gregory P. Baxter, Bohuslav Brauner, A. Debierne, A. Leduc, T. W. Richards, Frederick Soddy, and G. Urbain. 1923. "Report of the International Committee on Chemical Elements." *The Journal of the American Chemical Society* 45 (4): 867–74.

Bicchieri, Cristina. 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. New York, NY: Oxford University Press.

Choppin, Gregory, Jan-Olov Liljenzin, Jan Rydberg, and Christian Ekberg. 2013. *Radiochemistry and Nuclear Chemistry*. 4th edition. Elsevier Academic Press.

Hendry, Robin Findlay. 2012. In *Philosophy of Chemistry, Volume 6 - 1st Edition*, edited by Andrea Woody, Robin Findlay Henry, and Paul Needham, 1st edition, 6:256–69. Oxford: North Holland, Elsevier.

IUPAC. 2019. "Chemical Element." IUPAC Gold Book. 2019. http://goldbook.iupac.org/terms/view/C01022.

Kripke, S. 1980. *Naming and Necessity*. Harvard University Press.

Paneth, F. A. 1962a. "The Epistemological Status of the Chemical Concept of Element (I)." *The British Journal for the Philosophy of Science* 13 (49): 1–14.

———. 1962b. "The Epistemological Status of the Chemical Concept of Element (II)." *The British Journal for the Philosophy of Science* 13 (50): 144–60.

Putnam, H. 1975. "The Meaning of 'Meaning.'" In *Mind, Language and Reality*, 215–71. Cambridge: Cambridge University Press.

Scerri, Eric. 2000. “Realism, Reduction and the Intermediate Position.” In *Of Minds and Molecules*, edited by N. Bhushan and S. Rosenfeld, 51–72. Oxford: Oxford University Press.

———. 2007. *The Periodic Table: Its Story and Its Significance*. New York: Oxford University Press.

The "Inch-Worm Episode": Reconstituting the Phenomenon of Kinesin Motility

Introduction

Philosophical models of how phenomena are "reconstituted" in science tend to emphasize the importance of explanatory considerations in driving phenomenon reconstitution. On such models, phenomena are reconstituted as researchers gain insight into the explanatory mechanisms underpinning phenomena of interest (Bechtel and Richardson 1993/2010; Craver 2007), or as researchers recognize that their favored explanans is better suited to explain a phenomenon occurring at a "level of abstraction" higher than was initially assumed (Kronfeldner 2017).[1] This emphasis is perhaps unsurprising as mechanistic philosophy of science has, by and large, focused its efforts on *explanation* leaving the phenomena themselves construed as little more than the target thereof. That said, a number of philosophers following (Bogen and Woodward 1988) have considered the ways in which scientists treat phenomena as objects of investigation in their own right.[2] This paper follows in that tradition, analyzing a case of phenomenon reconstitution that occurred entirely within an experimental program dedicated to characterizing, rather than explaining, the phenomenon of kinesin movement.

Research on kinesin—a molecular motor that transports cargo around cells by moving unidirectionally along microtubule protofilaments—involves a substantial amount of experimental work dedicated to characterizing the phenomenon of kinesin movement. Unlike with macroscopic objects whose movements are readily observable, molecular motor movement

[1] Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press. Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT press. *philosophical perspectives on cognitive neuroscience*. Kronfeldner, M. (2015). Reconstituting phenomena. In *Recent Developments in the Philosophy of Science: EPSA13 Helsinki* (pp. 169-181). Springer, Cham.

[2] Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, *97*(3), 303-352.. Feest, U. (2011). What exactly is stabilized when phenomena are stabilized?. *Synthese*, *182*(1), 57-71. Colaço, D. (2018). Rip it up and start again: The rejection of a characterization of a phenomenon. *Studies in History and Philosophy of Science Part A*, *72*, 32-40.

is a phenomenon that takes place at the nanoscale. Characterizing it therefore presents challenges that require sophisticated experimental tools. In what follows, I focus on a particular tool, the *single-molecule motility assay*. Like patch-clamp recordings that made possible the characterization of the action potential and ion channels, the single-molecule motility assay enabled researchers to study the kinetic activities of single kinesin molecules and was an invaluable tool in the effort to characterize kinesin movement.

That the appropriate characterization of kinesin movement is that it walks "hand-over-hand" along microtubules was a guiding idea for researchers using the single-molecule motility assay.[3] In fact, the hypothesis was first suggested in 1989 in the very article reporting the development of this experimental tool. Over the following ten years, data from studies using variations on the basic design of the assay were interpreted as supporting hand-over-hand (HoH) walking, generating a limited consensus that, indeed, the correct characterization of the phenomenon of kinesin movement was that it walked HoH.

However, in 2002, a study involving a particularly interesting variation on this assay briefly disrupted this consensus, making a compelling case that kinesin walks in an "inch-worm" fashion rather than HoH. This study was quickly followed by a number of further single-molecule studies that re-established an even more robust HoH consensus. However, this is not a story of HoH advocates having been correct all along. Rather, the phenomenon of HoH walking was importantly "reconstituted" across the 2002 study.

In section I, I discuss the initial battery of single-molecule studies that were taken to support the HoH model of kinesin motility paying particular attention to the empirical criteria—

[3] This idea guided researchers using other methods as well, in particular, those using traditional biochemical techniques to study the hydrolytic cycle of the kinesin molecule. The interactions between the biochemical and single-molecule programs was important in the effort to map the stages of kinesin's mechanical steps to stages in its hydrolytic cycle. Here, I focus on the single-molecule program's attempts to characterize the molecule's mechanical steps.

processivity and *coordinated head activity*—that individuated HoH models as such and informed researchers' interpretations of their experimental results. Further, I describe the limitations this way of characterizing the phenomenon of HoH walking placed on the probative value of the single-molecule assay, leaving researchers to adjudicate between merely conceptually distinct HoH models with indirect, theoretical argumentation. Section II discusses an important 2002 study which exploited the latent experimental significance of ideas forwarded in the context of theoretical debate. This study re-drew the lines along which motility models were individuated, making *torque generation* the primary criterion. This new taxonomy enabled these researchers to design a more probative single-molecule study which lead them to reject HoH and forward an "inch-worm" model. Section III discusses the post-2002 studies that further exploited the new criterion for individuating motility models and secured consensus that kinesin walks hand-over-hand—now reconstituted as asymmetric HoH. Section IV concludes the article with a discussion of the case in light of extant philosophical models of phenomenon reconstitution.

As will be seen—and contrary to extant philosophical models—the reconstitution of kinesin motility did not occur in the context of attempting to *explain* the phenomenon, mechanistically or otherwise. Rather, it occurred entirely within the context of experimental efforts to characterize the phenomenon. More specifically, the reconstitution was driven by a recognition that individuating models of kinesin motility in terms of *torque generation* enhanced the probative value of the experimental program's primary investigative tool—the single-molecule motility assay. With this new taxonomy of motility models in hand, single-molecule researchers were able to use their assay to greater effect and establish a consensus that, indeed, kinesin walks hand-over-hand—now reconstituted as asymmetric hand-over-hand.

4

Section I: "Hand-Over-Hand" *circa* 1989 - 2002

By the 1980s, researchers had identified two molecules that function as motors – transforming energy into motion – myosin and dynein. In 1985, Vale and colleagues identified a third, kinesin, that was responsible for moving cargo such as organelles around the cell interior.[4]

Once kinesin had been identified and named, researchers turned to characterizing its structure and behavior. Bloom, Wagner, Pfister et al. (1988) subjected purified kinesin to centrifugation, differentiating two heavy and two light chains. They interpreted their results as showing that "bovine brain kinesin is a highly elongated, microtubule-activated ATPase comprising two subunits each of 124,000 and 64,000 daltons . . . and that the heavy chains are the ATP-binding subunits."[5] Electron microscope studies revealed globular heads at the N-terminal end of the heavy chains, which Scholey, Heuser, Yang et al. (1989) proposed serve both to bind to the microtubule and to be the locus of ATP hydrolysis.[6] They further hypothesized that the point of having two heads is that one remains attached to the microtubule while the other detaches and moves (Figure 1).

[4] Vale, R. D., Reese, T. S., & Sheetz, M. P. (1985). Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. *Cell*, *42*(1), 39-50.
[5] Bloom, G. S., Wagner, M. C., Pfister, K. K., & Brady, S. T. (1988). Native structure and physical properties of bovine brain kinesin and identification of the ATP-binding subunit polypeptide. *Biochemistry*, *27*(9), 3409-3416.
[6] Scholey, J. M., Heuser, J., Yang, J. T., & Goldstein, L. S. (1989). Identification of globular mechanochemical heads of kinesin. *Nature*, *338*(6213), 355.

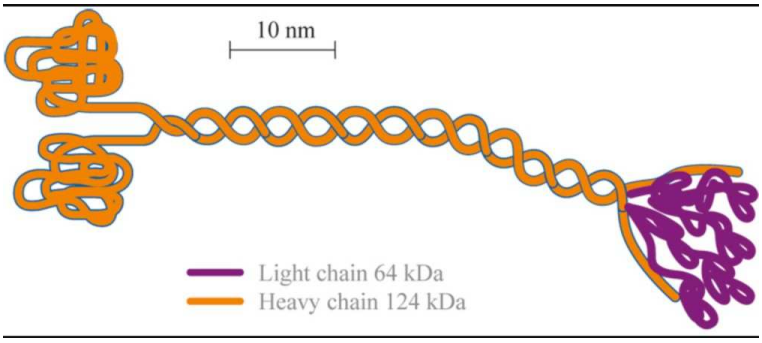


Figure 1: Kinesin molecule. The light chains (right) bind cargo and the heavy chains ("heads"; left) bind the molecule to the microtubule. The heads are also the site of ATP hydrolysis.

Howard, Hudspeth and Vale (1989) (henceforth, HH&V) reiterated this idea suggesting, on the basis of their findings using their newly developed technique for studying individual kinesin molecules, that it walks "hand-over-hand" along a microtubule. As their *single-molecule motility assay* became a central tool for investigating kinesin motility, it is worth explaining in some detail.

In order to develop an assay to investigate the motion produced by a single kinesin molecule, HH&V had first to establish that a single kinesin is capable of moving a microtubule in the first place. Their experimental design inverts how kinesin movement along microtubules may be normally understood—thinking of the microtubule as fixed and the kinesin as moving along it. Inverting this picture, these researchers immobilized kinesin molecules "heads-up" on glass cover slips in solutions containing progressively less kinesin to see how low they could go and still observe movement when microtubules were added. Their hypothesis was that if a single kinesin molecule could produce movement, they should observe microtubule movement at very low kinesin concentrations. Initially finding that only when kinesin density exceeded a rather high threshold did microtubules move, these researchers distinguished two hypotheses—first, that kinesin-induced microtubule movement is a highly collaborative affair requiring a number of

kinesin molecules working in concert and, second, that kinesin denatures when adsorbed onto the coverslips and only when a sufficient number of molecules are present do a few adsorbed kinesins remain in a conformation that can support movement. Clearly, the first hypothesis, if true, would be damning for the prospects of developing an assay meant to study movement produced by a single molecule.

Optimistically assuming the latter hypothesis, HH&V pre-treated the coverslips to prevent the hypothesized denaturation. Their optimism paid off. They found that they could produce microtubule movement with one-third of the kinesin concentration required with non-treated coverslips. The clincher, however, was the character of the microtubule movement that they observed:

> Each moving microtubule rotated erratically about a roughly vertical axis through a fixed point on the surface . . . presumably as a result of thermal forces, or of torques produced when a kinesin molecule bound to different protofilaments. When its trailing end reached this nodal point, the microtubule dissociated from the surface and diffused back into solution.[7]

The nodal point, these researchers concluded, was a single kinesin molecule. Thus, they found that a single kinesin, immobilized on a glass cover-slip, can move a microtubule and, at the same time, developed a technique for studying this movement that would prove central to the investigation of the phenomenon of kinesin motility.[8] More specifically, they found that a single kinesin can move a microtubule several micrometers. They reasoned that kinesin can remain

[7] Howard, J., Hudspeth, A. J., & Vale, R. D. (1989). Movement of microtubules by single kinesin molecules. *Nature*, *342*(6246), 154. Notice the mention of "torque." The idea that HoH walking may produce torque was on the table very early on. As we will see, however, this factor was thoroughly backgrounded in subsequent discussions of experimental results taken to bear on the HoH model of kinesin motility.

[8] Interestingly, they compare the probative force of their assay with that of patch-clamp recording designed to study the activity of single ion channels in neurons: "like patch-clamp recording from ion channels, the study of movement produced by single motor molecules provides an assay sensitive enough to monitor the activity of an individual protein molecule." *Ibid.*, 158.

attached to a microtubule by one of its heads, pushing the microtubule along as the other head moved forward, through 200 – 1000 iterations of its hydrolytic cycle. Linking this finding to the fact that the molecule has two globular heads, these researchers suggested that the molecule works "hand-over-hand" with one head always remaining attached to the microtubule. However, they also suggest an alternative possibility. Here is the full quote:

> It is possible that kinesin's two globular heads work hand-over-hand, so that *one head is always bound* and prevents the microtubule from diffusing away. Alternatively, the *two heads may work independently* . . . If this is so, the time in the reaction cycle during which the kinesin heads are detached from the microtubule must be so brief, probably less than 1 ms, that the microtubule is unlikely to diffuse out of reach of the kinesin molecule (my emphasis).[9]

It's important to attend closely to what "hand-over-hand" meant from the point of view of this 1989 experiment. The contrast HH&V draw between their alternatives makes clear that, as opposed to a model on which the heads *work independently* and, thus, on which the whole molecule (both heads) detaches from the microtubule, the "hand-over-hand" model has it that the kinesin heads *coordinate* their activity such that the molecule remains attached to the MT by at least one head during its walk. In other words, HoH walking consists in 1) the molecule remaining attached to the MT (*processivity*) by at least one head by means of 2) *coordinated head activity*. These became the empirical criteria that were taken by subsequent researchers to individuate HoH models as such and which informed the interpretation of experimental results for the next decade.

Over the course of the following decade, two versions of the single-molecule assay developed. 1) "MT-gliding assays" in which kinesin molecules are immobilized to glass cover

[9] *Ibid.*, 158

slips and microtubule movement is observed and 2) "bead assays" in which microtubules are immobilized and kinesin-bound beads are observed to move as the kinesin attaches to and walks along the immobilized microtubule. Both "geometries" of the single-molecule assay lent support to both aspects of HH&V's HoH hypothesis.

Not all studies were immediately univocal in this respect, however. In a version of the bead assay, Block, Goldstein and Schnapp (1990) immobilized microtubules, rather than kinesin, on glass cover-slips. Coating silica beads with carrier protein and exposing them to low concentrations of kinesin, these researchers were able to observe the beads as single kinesin molecules moved them along the immobilized microtubule tracks. Using optical tweezers—which split laser beams to trap kinesins—to individually manipulate the moving beads, they found that under the forces exerted by the optical trap, the bead would detach from the microtubule after, on average, 1.4 μm and be pulled back toward the center of the trap.[10] This, they argued, provides support for the claim that, "the kinesin molecule might detach briefly from the substrate during each mechanochemical cycle" (not processive) and referred to their alternative model of kinesin motility as a "stroke-release" model.[11]

However, a number of influential single-molecule studies over the next 10 years strongly supported HoH over the non-processive stroke-release model. In a clever variation on the MT-

[10] The invention of optical tweezers was significant for research on kinesin motility in ways beyond those discussed here. For instance, since kinesin motility is a phenomenon occurring at the nano-scale, thermal forces are relevant. It is therefore difficult to discern what observed motion is Brownian motion and what is due to the action of the molecule. Having kinesin move cargo against the forces exerted on it by the "trap" ensures that whatever motion is observed is due to the molecule's action. This technique enabled Svoboda, Schmidt, Schnapp et al. (1993) to observe abrupt transitions of 8 nm steps, a distance that corresponds to the repeat distance between successive α-β tubulin dimers. They propose "that the two heads of a kinesin molecule walk along a single protofilament—or walk side-by-side on two adjacent protofilaments—stepping ~8 nm at a time, making one step per hydrolysis (or perhaps fewer, requiring multiple hydrolyses per step)" Svoboda, K., Schmidt, C. F., Schnapp, B. J., & Block, S. M. (1993). Direct observation of kinesin stepping by optical trapping interferometry. *Nature*, *365*(6448), 721.

[11] Block, S. M., Goldstein, L. S., & Schnapp, B. J. (1990). Bead movement by single kinesin molecules studied with optical tweezers. *Nature*, *348*(6299), 348. These researchers also suggested a model on which the molecule is always bound by at least one head but "weakly" – just strong enough to remain attached in the face of thermal forces, but not strongly enough to remain attached when subjected to the forces of the optical trap.

gliding assay, Ray et al. (1993) constructed microtubules consisting of 12, 13 or 14 protofilaments (12-mers, 13-mers, 14-mers). Protofilaments of 13-mers run parallel to the MT axis while 12 and 14-mers exhibit right- and left-handed helical organizations ("twists") respectively. Observing the movement of these microtubules induced by single immobilized kinesin molecules, the researchers found that the 12 and 14-mers rotated with the pitch and handedness predicted by the hypothesis that the kinesin molecule follows the protofilament axis. That kinesin movement is constrained in this way—that it "tracks the protofilament"—suggested that at least one head remains attached to the MT during its walk, therefore lending support to that aspect of the HoH model of kinesin movement.[12]

In a version of the bead assay, Berliner et al. (1995) attached single-headed kinesin derivatives to streptavidin-coated polystyrene beads and found that, unlike intact kinesin or two-headed constructs, the single-headed molecule moved beads perpendicular with respect to the microtubule axis and failed to drive continuous unidirectional movement. This perpendicular movement suggested that the single-headed molecules lack the means to maintain their association with a particular protofilament track, namely, another head with which to coordinate its activity. The absence of perpendicular movement suggested that the opposite is true for two-headed kinesin, lending support to the idea that the activity of the two heads is coordinated to ensure that one head remains MT-bound at all times. This, in turn assures that the molecule tracks the protofilament axis as it was found to do in the study described in the paragraph above.[13]

Further support for the HoH model came with the introduction of fluorescent labelling in the single-molecule assay. In a version of the MT-gliding assay, Vale et al. (1996) directly

[12] Ray, S., Meyhöfer, E., Milligan, R. A., & Howard, J. (1993). Kinesin follows the microtubule's protofilament axis. *The Journal of cell biology*, *121*(5), 1083-1093.

[13] Berliner, Elise, Edgar C. Young, Karin Anderson, Hansraj K. Mahtani, and Jeff Gelles. "Failure of a single-headed kinesin to track parallel to microtubule protofilaments." *Nature* 373, no. 6516 (1995): 718-721.

observed the movement of individual fluorescently labeled kinesin molecules finding that the labeled two-headed kinesin travels an average distance of 600nm per encounter with a microtubule whereas single-headed constructs shows no detectable movement.[14] This corroborated Berliner et al. (1995)'s finding discussed above, suggesting that the two heads working together is required for movement.

Hancock and Howard (1998) immobilized single-headed kinesin onto glass cover slips and found that a minimum of four to six single headed molecules are necessary to produce movement. They further showed that, even at high ATP concentration, the single-headed molecules detached from microtubules 100-fold more slowly than their two-headed counterparts "directly support[ing] a coordinated, hand-over-hand model in which the rapid detachment of one head . . . is contingent on the binding of the second head."[15] Thus, their study demonstrated a degree of "chemical coordination" between the two heads lending biochemical substance to the idea that kinesin motility involves coordinated head activity.

While single-molecule studies such as these generated a limited consensus that kinesin walks HoH, a number of motility models that met the HoH criteria and were consistent with extent single-molecule data were *conceptually* distinguished in the literature during this time. However, without empirical criteria by which to distinguish them *experimentally*, it was left to single-molecule researchers to adjudicate between these models by way of indirect argumentation that appealed to data from sources external to the single-molecule program.

To illustrate, (Figure 3) on page 13 distinguishes five stepping patterns understood to be variably consistent with the data to that time. Findings regarding the structure and dimensions of the molecule, the lattice structure of microtubules and the sites on tubulin heterodimers to which

[14] Vale, Ronald D., Takashi Funatsu, Daniel W. Pierce, Laura Romberg, Yoshie Harada, and Toshio Yanagida. "Direct observation of single kinesin molecules moving along microtubules." *Nature* 380, no. 6573 (1996): 451-453.
[15] Hancock, W. O., & Howard, J. (1998). Processivity of the motor protein kinesin requires two heads. *The Journal of cell biology*, *140*(6), 1395.

kinesin was understood to bind provided fodder for indirect arguments in favor of or against such conceptually distinguished models. (see Cross, 1995; Howard, 1996; Block, 1998 for reviews).[16] As we see in (Figure 2), microtubules consist in protofilaments arranged in cylindrical fashion. Each protofilament consists of alternating tubulin (α- and β-tubulin) heterodimers.

[16] Cross, R. A. (1995). On the hand over hand footsteps of kinesin heads. *Journal of muscle research and cell motility*, *16*(2), 91-94. Howard, J. (1996). The movement of kinesin along microtubules. *Annual review of physiology*, *58*(1), 703-729. Block, S. M. (1998). Kinesin: what gives?. *Cell*, *93*(1), 5-8. For micrographic data relevant to these indirect arguments see: Kikkawa, M., Ishikawa, T., Nakata, T., Wakabayashi, T., & Hirokawa, N. (1994). Direct visualization of the microtubule lattice seam both in vitro and in vivo. *The Journal of cell biology*, *127*(6), 1965-1971. Song, Y. H., & Mandelkow, E. (1995). The anatomy of flagellar microtubules: polarity, seam, junctions, and lattice. *The Journal of cell biology*, *128*(1), 81-94. Harrison, B. C., Marchese-Ragona, S. P., Gilbert, S. P., Cheng, N., Steven, A. C., & Johnson, K. A. (1993). Decoration of the microtubule surface by one kinesin head per tubulin heterodimer. *Nature*, *362*(6415), 73.

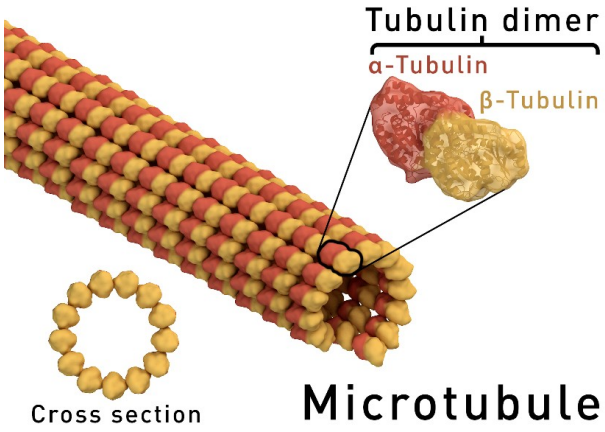


Figure 2: Microtubule structure.

Several biochemical studies suggested that a tubulin heterodimer can bind only one kinesin head (Song and Mandelow, 1993; Walker, 1995; Tucker and Goldstein 1997). This fact, coming from outside the single-molecule program, was appealed to in adjudicating between conceptually distinct models. For instance, as we see in (Figure 3), an "inchworm model" had been distinguished prior to 2002. On this model, one head always remains in the lead with the other head trailing behind.[17] This model, however, requires each tubulin dimer to have two binding sites (or a single, shared binding site) so that the two heads could be brought into proximity with one another. This, argued Block and Svaboda (1995), was difficult to square with binding patterns gleaned from the aforementioned biochemical studies. They note further that such a model involves an implausibly more complicated step consisting of a "two-part cycle comprising

[17] Though not a "hand-over-hand" model in what is perhaps the intuitive sense of the phrase, by the lights of the empirical criteria that distinguished HoH models as such (distinguished them from e.g. stroke-release models) "inchworm" models were a species of HoH. As we will see, it was not until the introduction of a new empirical criterion that inchworm models were adequately distinguished from HoH models along empirically tractable lines.

the successive action of both heads."[18] That is, rather than each 8nm step consisting of a single head relocating to the next tubulin binding site, it would involve, first, the lead head moving and, second, the trailing head moving up from behind to keep pace.

These same researchers also argued that "long stride" seemed implausible on the grounds that it required the relatively small kinesin molecule to extend a full 16nm to move the centroid of the molecule 8nm as had been observed in their motility assays. Since this would require that the stalk connecting kinesin's heads be capable of this kind of extension, Long Stride was deemed speculatively possible at best. Cross (1995) seems to have the same worry in mind in criticizing motility models that require kinesin to stretch its heads across a protofilament, straddling it on either side, and walking along the protofilaments adjacent to it. This would be like "two-step I" only with the squares moved over one protofilament to the right. Cross says of such a model that it is "barely credible."[19]

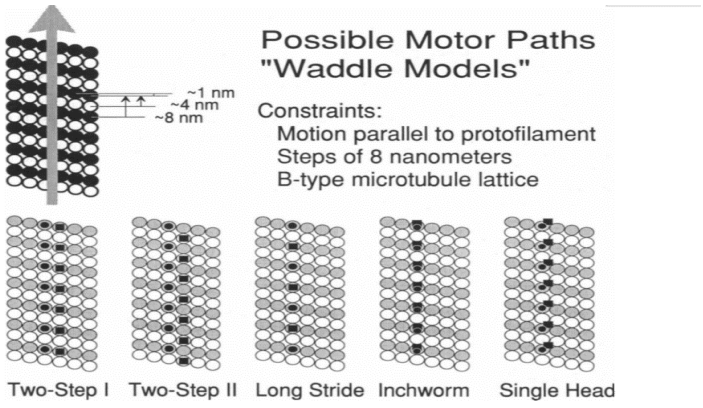


Figure 3: Conceptually distinguished motility models the plausibility of which was left to be adjudicated by indirect arguments based on data coming from outside the single-molecule program.

[18] Block, S. M., & Svoboda, K. (1995). Analysis of high resolution recordings of motor movement. *Biophysical journal*, *68*(4 Suppl), 237s.
[19] Cross, R. A. (1995). On the hand over hand footsteps of kinesin heads. *Journal of muscle research and cell motility*, *16*(2), 92.

14

This kind of indirect argumentation was characteristic of attempts to adjudicate between the motility models that had been conceptually distinguished in the first ten years of single-molecule research. While most researchers agreed that HoH was the correct characterization of kinesin motility (rather than "stroke-release"), a number of HoH models could be distinguished that were consistent with single-molecule data. Thus, a space of merely conceptually distinct models existed to which researchers using the single-molecule motility assay had no experimental access. They were therefore left with indirect argumentation based on findings from experimental sources external to the single-molecule research program.

Notably absent from most of this indirect argumentation were considerations of *torque*. This, despite the fact that HH&V had mentioned it in the very paper in which they coined the phrase "hand-over-hand." There was an exception, however. In an impressively comprehensive review, Howard (1996) did bring the idea that HoH walking produces torque into the discussion along with a number of other considerations the experimental significance of which would be exploited in a 2002 study that represented a significant challenge to the hand-over-hand consensus.[20]

Howard (1996)'s indirect argument represents a compelling theoretical analysis. He assumes, on the basis of analogy with other known molecular motors, that kinesin has a "two-fold axis of rotational symmetry" and infers that, therefore, the heads are functionally equivalent – "they have the same hydrolysis cycles and make the same motions."[21] He calls this the "equivalence hypothesis." Tracing out the consequences of this hypothesis in conjunction with extant experimental data, Howard argued that the most plausible model for kinesin motility was

[20] Howard, J. (1996). The movement of kinesin along microtubules. *Annual review of physiology*, *58*(1), pp. 724.
[21] For an intuitive sense of what having a "2-fold axis of rotational symmetry" means, imagine two chairs facing each other on either side of a line and equidistant from that line. Rotating one chair 180 degrees with respect to that line will bring that chair into the precise position of its mate. Howard assumed that the relation between kinesin's two heads was the same.

a “rotary model” on which the molecule’s heads pass each other on the same side each step (Figure 3) rather than on alternating sides like the way in which our human legs move past each other as we walk.

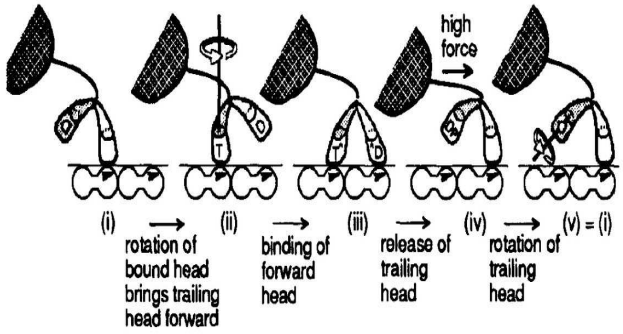


Figure 3: Each head has the same hydrolysis cycle and the same stepping movement, the stepping head always passing the MT-bound head on the same side. Notice that state (i) is identical to state (v).

His argument involves three key ideas the experimental significance of which was only realized later. First, taking his equivalence hypothesis in conjunction with the protofilament tracking data discussed above, Howard argues against models like the ones labeled **Two-Step** in figure 1. According to such models, the molecule switches back and forth, alternately binding adjacent protofilaments with each head. Assuming the equivalence hypothesis, a consequence of which is that the beginning of each step finds the molecule in the same 3D conformation, Howard argues that if one head, attached to a protofilament (a) were to undergo a conformational change and motion so as to bring the other head to an adjacent protofilament (b), then the equivalent conformational change in head 2 - required by the equivalence hypothesis - would bring head 1 to the next protofilament (c). This would induce a rotation in the 13-mer microtubules that was not observed in the single-molecule study discussed above. *Inter alia,* this reasoning leads Howard to his rotary model. As for the second key idea, Howard notes a “seemingly

unthinkable" consequence of this model. Because of the assumed equivalence between the heads, the molecule will always rotate in the same direction and "Thus the tail (and organelle) will tend to wind up like the rubber band of a toy airplane."[22] Howard suggests that this torsion could be accommodated by the torsional flexibility the neck was found to exhibit in an earlier study (Hunt and Howard 1993).[23] That the neck has this torsional flexibility is the third key idea.

The experimental significance of these three ideas—1) the equivalence hypothesis, 2) that kinesin motility may produce torque which is communicated to the cargo and 3) that the kinesin neck is torsionally flexible—later came to be appreciated and exploited in a study that introduced a new empirical criterion for individuating motility models. Recall, from the late 1980s to the late 1990s, the criteria that individuated HoH models as such were that 1) the molecule is genuinely *processive* and that it is so by means of 2) *coordinated head activity*. From the point of view of this taxonomy, a number of HoH motility models could be conceptually distinguished that were more or less consistent with available experimental data but adjudicating between them was left a matter of indirect argumentation using data from sources external to the single-molecule program. As we'll see, Hua et al.'s 2002 study re-drew the taxonomic lines and, as a result, lent further probative value to the single-molecule motility assay.

[22] Howard, J. (1996). The movement of kinesin along microtubules. *Annual review of physiology*, *58*(1), pp. 724.
[23] Hunt, A. J., & Howard, J. (1993). Kinesin swivels to permit microtubule movement in any direction. *Proceedings of the National Academy of Sciences*, *90*(24), 11653-11657.

Section III: Hand-over-Hand vs. Inchworm

Hua, Chung, and Gelles (2002) inaugurated an important shift in the empirical criteria by which motility models were individuated.[24] As mentioned above, their study exploited ideas that had been floated in the literature in the context of indirect, theoretical argumentation. First, the design of the experiment was a modified version of (Hunt and Howard 1993)'s assay used to measure the torsional flexibility of the kinesin neck. However, rather than using native kinesin which, in that study, had been found to have a *flexible* neck, Hua and colleagues used a *stiff-necked,* two-headed biotinated kinesin derivative (K448-BIO). This ensured that the connection between the microtubule, this molecule, and the glass cover slip on which the molecule was immobilized would be torsionally stiff, thus guaranteeing that if torque was indeed generated by the walking molecule, as Howard's model predicted, it would not be taken up by a flexible neck. Rather, it would be communicated to the cargo and generate a clearly observable 180-degree rotation of the microtubule with each step of the molecule. Their design, therefore, took the "seemingly unthinkable" consequence Howard had traced out eight years earlier and cleverly turned it into an intervention.

Further, they pointed out that whether the heads of the molecule pass each other on the same side, as in Howard's rotary model, or pass each other on alternating sides, the orientation of the molecule relative to the microtubule axis would switch as the heads alternate between being the leader and being the follower. This, in turn, would generate torque, and induce an observable microtubule rotation. In other words, the differences between the *intermediate* states of rotary models and left-right alternate stepping models were immaterial. What mattered for torque

[24] Hua, W., Chung, J., & Gelles, J. (2002). Distinguishing inchworm and hand-over-hand processive kinesin movement by neck rotation measurements. *Science*, *295*(5556), 844-848.

generation was that the molecule *begins* each step in the same 3D conformation only with the heads swapping between leading and following. Hua et al., dubbed these torque generating models *symmetric hand-over-hand* (Figure 3A). By the lights of the criterion of torque generation, both Howard's rotary model and alternate left-right stepping models count as symmetric HoH models.

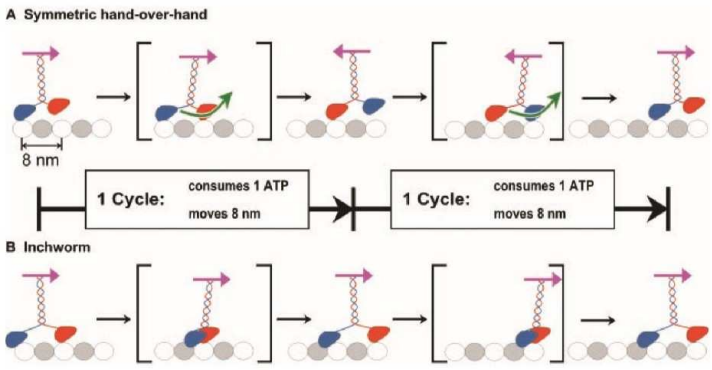


Figure 3: The brackets around the intermediate stages of the steps in A indicate their irrelevance. Whether the stepping head passes on the same side, as it does in the diagram, or passes on alternating sides of the bound head, the molecule will change its orientation as indicated by the arrows on top of the molecule.

To appreciate the shift in criteria for individuating motility models these researchers introduced, consider the sense in which Howard's rotary model would be considered a species of HoH model prior to this study. It would count as an HoH model because it sees the molecule as remaining attached to the microtubule by at least one head (processivity) and that it does so by means of coordinated head activity. The same goes for alternate left-right stepping models. From the point of view of the new criterion—torque generation—both count as HoH models but for very different reasons. First off, they would no longer count as HoH models full stop. Rather they would be considered instances of *symmetric* HoH to be distinguished from an *asymmetric*

HoH model—a distinction I will discuss in more detail shortly. Further, rather than processivity or coordinated head activity serving to distinguish them as HoH models (as opposed to stroke-release), they count as (symmetric) HoH models because they generate torque. This, again, for the reason that both models view the molecule as beginning each step in the same 3D conformation, rotating its orientation relative to the microtubule axis during its step and, thus, generating torque.

It was with respect to torque generation that the distinction between symmetric HoH and asymmetric HoH was drawn. Asymmetric HoH models deny that the molecule generates torque by denying the equivalence of the heads' steps. On this model, kinesin alternates between *two distinct conformations*—a different one at the beginning of each step—"in precisely such a way as to cancel the 180-degree reorientation induced by head alternation."[25]

Finally, and most importantly, after this re-drawing of the taxonomic lines, "inchworm" was no longer to be considered a sub-species of HoH as it was by the lights of the pre-2002 empirical criteria—processivity and coordinated head activity. Now, with torque generation serving to individuate models, inchworm was distinguished from HoH along empirically tractable lines.

Armed with this more probative empirical criterion by which to individuate motility models, Hua et al. (2002) developed and ran their single-molecule assay, failing to observe the microtubule rotations predicted by symmetric HoH models. They therefore rejected that characterization of the phenomenon of kinesin motility. This left two non-torque generating possibilities: 1) that the molecule walks in an asymmetric HoH fashion or 2) that it walks inchworm-style. In a way reminiscent of the indirect arguments discussed above, Hua and colleagues argued against the plausibility of asymmetric HoH. In brief, they found it implausible

[25] Hua et al. 847.

that the differences between 3D conformations at the start of each step could be such that they could exactly compensate for the rotation and, in turn, the torque produced by an asymmetric walk. Rejecting asymmetric HoH on these grounds, these researchers argued that the correct characterization of the phenomenon of kinesin motility is that it walks in an "inchworm" fashion.

So, what led these researchers to reject HoH as an appropriate characterization of the phenomenon and adopt inchworm? Note that although their rejection is experimentally motivated, they did not experiment for the purpose of gathering evidence to undermine that which had already been found in support of the HoH model. That is, they did not gather evidence to undermine the single-molecule studies that had supported the claim that the molecule is processive and that its heads coordinate their activity. Thus, they did not employ a "defeater-strategy" as in the case of "memory transfer" discussed by Colaco (2019). Rather, as described above, they recognized the experimental significance latent in certain ideas that had already been floated in the literature. They then constructed a new taxonomy using torque generation as the criterion for individuating motility models which, in turn, enabled them to design a more probative version of the single-molecule motility assay. It further enabled them to recognize an important distinction—that between *symmetric* and *asymmetric* HoH models. Their single-molecule study, they recognized, only bore directly on symmetric HoH models. Their study refuted symmetric HoH leaving the refutation of the asymmetric model to be done by indirect argumentation. Thus, between their empirical results and indirect argumentation, they rejected symmetric and asymmetric HoH models respectively, and defended inchworm as the most plausible model for the phenomenon of kinesin motility.

Section IV: Further Experimental Implications of the New Taxonomy

In section I, we noted the role that indirect argumentation played in adjudicating between conceptually distinct models. While such arguments, in addition to the single-molecule data, led to a limited consensus, they were not decisive in adjudicating between available HoH motility models. However, these more theoretical arguments led to ideas that had latent experimental significance. It was just a matter of unlocking it. The empirical criteria in terms of which models of kinesin motility were initially individuated— processivity and coordinated head activity—left open an experimental dead-space seemingly inaccessible to the single-molecule assay. The key granting the single-molecule assay experimental access to the dead-space was torque generation. Turning this key generated a new taxonomy, one enabling the development of a more probative variation of the single-molecule motility assay.

The studies that emerged in the following two years took advantage of this more experimentally tractable taxonomy, re-securing a consensus that kinesin walks HoH—now reconstituted as asymmetric HoH. Kaseda et al. (2003) tested the inchworm model's prediction that only one head is hydrolytically active. These researchers used optical tweezers in a bead assay to measure the stepping rate of kinesins mutated such that one head hydrolyzes ATP more slowly than the other. If both heads are hydrolytically active, they reasoned, their mutant molecule should show a "limp" in its stepping pattern as it walks. This is in fact what they observed undermining the inchworm models prediction of single-head catalysis.[26] That same year, Asbury et al. (2003), using optical tweezers in a bead assay, found that kinesin constructs with two identical wild-type heads also show a "limp" in their stepping suggesting that the

[26] Kaseda, K., Higuchi, H., & Hirose, K. (2003). Alternate fast and slow stepping of a heterodimeric kinesin molecule. *Nature Cell Biology*, *5*(12), 1079.

molecule alternates between two conformations from step to step thus supporting asymmetric HoH walking.[27] Yildez et al. (2004) directly observed the movement of kinesin heads tagged with a fluorescent dye and found that each head moves 16nm per step and also that the tagged heads pause after each movement presumably while the other untagged head moved. These findings are inconsistent with the inchworm model which takes each head to move 8nm per ATPase cycle and supports an asymmetric HoH model.[28] Higuchi et al. (2004) observed a difference in the timing of every other step in kinesins with identical mutations in the nucleotide-binding sites in each head.[29] The limping they observed is similar to that observed by Asbury and colleagues above, but more pronounced due to the mutation.

Each of these studies exploited the reimagined taxonomy of motility models inaugurated by Hua et al. (2002). Interestingly, it was no advancement in tool-development that enabled researchers to observe kinesin's "limping" step. The instrumentation necessary to do so—the single-molecule bead assay and optical tweezers—had been in place for over a full decade prior to its being observed. It was rather a conceptual innovation ushered in by the new taxonomy that enabled researchers to look for kinesin's limping step and appreciate its significance. In fact, even if the limping step had been observed prior to this reconstitution of the phenomenon, it is not obvious that researchers would have recognized its significance, at least not in the way that it was recognized afterwards. It was in observing kinesin's limp against the backdrop of a taxonomy of motility models which included the category of asymmetric HoH that its significance for experimental work in characterizing the phenomenon of kinesin motility became apparent. Therefore, although recent philosophical efforts to emphasize innovative tool-

[27] Asbury, C. L., Fehr, A. N., & Block, S. M. (2003). Kinesin moves by an asymmetric hand-over-hand mechanism. *Science*, *302*(5653), 2130-2134.
[28] Yildiz, A., Tomishige, M., Vale, R. D., & Selvin, P. R. (2004). Kinesin walks hand-over-hand. *Science*, *303*(5658), 676-678.
[29] Higuchi, H., Bronner, C. E., Park, H. W., & Endow, S. A. (2004). Rapid double 8-nm steps by a kinesin mutant. *The EMBO journal*, *23*(15), 2993-2999.

development in driving scientific research are to be applauded, the case of the "inch-worm episode" reminds us conceptual innovation remains an important factor.[30]

Section V: The "Reconstitution" of Hand-over-Hand Walking

As I mentioned in my introduction, and as the history I have laid out reveals, the story of the re-establishment of the HoH consensus is not one according to which HoH advocates were shown to have been right all along. Rather, the phenomenon of HoH walking was importantly reconstituted across the inchworm episode from HoH to asymmetric HoH. The inchworm episode and the reconstitution it inaugurated took place entirely within the context of an experimental program dedicated to characterizing, rather than explaining, the phenomenon of kinesin motility. This is of particular philosophical interest as standard philosophical models of phenomenon reconstitution have it that explanatory considerations drive phenomenon reconstitution.

Bechtel and Richardson (1993/2010)'s model of phenomenon reconstitution, for instance, was motivated by their case study of the "Mendelian trait."[31] Classically, the Mendelian trait was understood as a macroscopically observable phenotypic trait. Faced with the fact that patterns of phenotypic inheritance could not be explained in terms of single genes – "phenotypic traits were the products of many genes in a complex organization"—researchers in the middle of the 20th century abandoned the phenotypic trait as the central Mendelian unit in favor of a unit at a lower level of mechanistic analysis, the *enzyme*. Thus, the explanandum phenomenon to be accounted for in terms of single genes was reconstituted, shifting it down from the phenotypic trait to the enzyme, in the effort to develop mechanistic accounts of gene action.

[30] Bickle, J. (2016). Revolutions in neuroscience: Tool development. *Frontiers in systems neuroscience*, *10*, 24.
[31] Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT press.

Craver (2007) discusses a further way in which phenomena can be reconstituted in the context of seeking mechanistic explanations. According to Craver, phenomena can be reconstituted in the wake of researchers recognizing that they have committed one of two errors – the "lumping error" or the "splitting error."[32] Both errors require inquiry into the phenomenon to have developed to a point at which researchers have both a characterization of the phenomenon and putative mechanistic explanations on the table. Scientists observe they have committed the splitting error when they recognize that they have erroneously thought that some phenomena of interest are due to two or more distinct types of mechanisms when, in fact, they are due to mechanisms of the same type. They may then reconstitute the phenomena such that where once they thought of them as two distinct phenomena underpinned by two distinct types of mechanisms, they now understand them as one phenomenon underwritten by a single mechanism-type. The lumping error, on the other hand, occurs when a particular phenomenon is thought to be generated by a single mechanism while, in fact, two distinct mechanisms underwrite the phenomenon. In light of recognizing this error, scientists may reconstitute the phenomenon, considering it now as two distinct phenomena.

(Kronfeldner 2015)'s model differs from both of the above. She describes how phenomenon reconstitution can result not only as a result of researchers gaining insight at the level of mechanism, but also by researchers "moving up to a level of greater abstraction." [33] To illustrate, a researcher interested in explaining a particular phenotypic trait of a particular person - their height, say - will be unable to do so as it is widely recognized that such traits are the result of complex interactions between an individual's genetic inheritance and their ontogenetic environment. This does not mean, however, that genes do not explain. By moving up to an

[32] Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press. pp. 123-124.
[33] Kronfeldner, M. (2015). Reconstituting phenomena. In *Recent Developments in the Philosophy of Science: EPSA13 Helsinki* (pp. 169-181). Springer, Cham.

explanandum phenomenon at a greater level of abstraction, e.g. average differences between the heights of males and females in a population, researchers can appeal explanatorily to differences in genotype, ignoring the complexity introduced by gene-environment interactions. In this way, researchers can hold fast to a particular "causal factor" in terms of which they wish to pitch their explanations and constitute the phenomena to be explained accordingly.

All three models have it that phenomenon reconstitution is driven by explanatory considerations. The research on kinesin motility discussed throughout this paper, however, involves experimental work dedicated solely to characterizing the phenomenon of kinesin movement. Developing mechanistic explanations of kinesin movement (not discussed) involves researchers determining how the energy released from ATP-hydrolysis occurring in the molecule's nucleotide binding sites results in structural changes throughout the molecule. Mechanistic explanation asks after the role played (if any) by thermal forces in bringing the heads forward in their stepping pattern. It attempts to determine whether elastic tension on the neck linker generated as the molecule stretches during its walk provides energy—in addition to that provided by ATP-hydrolysis—that may or may not be necessary for walking.[34] These (and further issues) are, of course, important for developing mechanistic explanations for kinesin motility—for answering the question of *how* kinesin manages to walk in the way it does. But considerations at this explanatory level did not, as we saw, figure into the reconstitution story. Again, it took place entirely within the context of experimental efforts to characterize the phenomenon—to characterize the *way kinesin walks*, not *the means by which* it manages to walk that way.

In closing, Colaco (2020) notes "there is a lacuna in the literature regarding how researchers determine whether their characterization of a target phenomenon is appropriate for

[34] Ref to Bechtel and Bollhagen "Molecular Motors: Transforming energy to motion."

their aims."[35] This paper helps to illuminate that lacuna. In order to experimentally adjudicate between alternative characterizations of kinesin motility, single-molecule researchers sought *empirical* criteria by which to individuate them—criteria that distinguished them along lines that were testable from the point of view of the single-molecule motility assay. It was determined that individuating models of kinesin by appeal to torque generation rather than merely processivity and coordinated head activity, enabled access to what was antecedently an experimental dead-space consisting of merely conceptually distinct motility models. The new taxonomy rendered that space experimentally accessible to the single-molecule assay. Thus, the "inchworm" episode illustrates how researchers can recharacterize phenomena to the end of enhancing the probative value of their experimental tools.

[35] Colaço, D. Recharacterizing scientific phenomena. *Euro Jnl Phil Sci* **10,** 14 (2020). https://doi.org/10.1007/s13194-020-0279-z

27

Growing Knowledge: Epistemic Objects in Agricultural Extension Work

Abstract: We outline a specialized form of knowledge arising from established communication practices between farmers, university researchers, and regulators. The *grower standard* is a benchmark concept in agricultural experiments that differs from familiar epistemic objects in philosophy of experiment such as controls or background conditions. It is a unique, institutionally-structured way in which agricultural experiments are value-laden. Grower standard is not a one-size-fits-all standard. It is the product of active interactions between diverse agricultural communities of stakeholders within *agricultural extension* communication practices. Exploring this form of knowledge coproduction, we explore the role extension work plays in shaping agricultural science more broadly.

1. Introduction

In Kentucky, agricultural experiments on tobacco crops need to be planted by June 20. Fungicide experiments on grapevines in Oregon begin when the plants achieve six inches of growth. In Missouri, cotton pest management experiments count the nodes above the highest first position of white flower (NAWF) to determine when to terminate insect control practices. These peculiarities of experimental design each originate from the concept of *grower standard*. Grower standard is a benchmark concept used in agricultural science. It furnishes the basis for comparison between farming practices and agricultural experiments.

Some considerations relevant to grower standard are similar to control conditions or background conditions discussed in the design of experiments in the natural sciences, while other considerations are wholly unfamiliar or are similar to considerations from the social sciences rather than the natural sciences. For instance, insecticide is used on cotton plants to minimize insect interference with developing cotton bolls. NAWF measures the flowering date of the last bolls. Once the cotton plant stops producing bolls, insect control ceases to make an economic difference in the overall yield. On the other hand, Kentucky tobacco experiments need to be planted by June 20 because that is the latest date that commercial growers can plant tobacco and be guaranteed insurance on their crops. Only experiments performed prior to that date will provide useful information to growers as farmers are well aware that growing conditions following the June 20 cutoff are substantially different than those prior to it.

In this paper, we characterize grower standard as an epistemic object of agricultural science and use this characterization to illustrate a unique and institutionally-structured way in which agricultural experiments are value-laden. In Section 2, we define grower standard and argue that it differs from familiar epistemic objects in philosophy of science. In Section 3, we show that one important reason that grower standard differs from these more familiar epistemic objects is that grower standard is a product of interactions between research communities and agricultural extension workers. We explore the role extension work plays in the shaping of agricultural science practices more broadly and describe this role in terms of knowledge coproduction. Section 4 concludes.

1.1 Agriculture: a glossary

Agriculture remains an area of research less familiar to philosophy of science (but cf. Thompson, 2017). In order to help our readers navigate this new area, we begin with a brief glossary.

- *Agricultural practice*: the stewardship of crops and livestock.
- *Agricultural sciences*: studies of the cultivation of soil for the growing of crops, husbandry of animals, management of land systems, global and local seed economics, food scarcity, biofuels, and more.
- *Agronomy*: the scientific study of crops, soil, and plant ecology. Its focus is on crops of high commercial value for food, fuel, or fiber.[1]
- A*gricultural extension*: a formalized system of communication practices established between farmers, university researchers, and regulators to exchange ideas about new agricultural research, technologies, and practices.
- *Agricultural extension work*: activities that include digital and on-site consulting, attending local and regional grower/producer meetings, giving field day presentations, and carrying out experiments to improve agricultural practices. By participating in these activities, extension workers and farmers exchange information about how to improve production, increase crop diversity, provide nutrient support to soil, manage irrigation practices, and control pests and diseases.

[1] Agronomy is primarily informed by biological and ecological considerations and methods, its close connection to agricultural practices means that it is deeply entangled with technological, economic, commercial, and sociopolitical concerns. This entanglement is a motivating reason for our present interest.

- *Agricultural extension specialist*: an academic researcher[2] whose professional duties include both the production of scholarship and the performance of extension work (alongside teaching and service). While research and extension work are evaluated as separate categories, Extension specialists usually perform both types of work on a single research domain.

2. Grower Standard as Epistemic Object

2.1 Defining Grower Standard

An important component of designing agronomy research protocol is to identify and recreate what is known as “grower standard,” sometimes referred to as “grower standard practice” or “standard grower conditions.” Conditions that specify a grower standard can include fungicide, herbicide, and insecticide protocols; fertilization, watering, and harvest methods and timing; soil treatments; instrumentation used (e.g., cotton-picker, transplanter, tiller); and pathogen containment strategies. What counts as grower standard for a given experiment is particular to the crop, region, scale of production, and type of farming practice (e.g., organic v. conventional).

“Grower standard” is regularly referenced in descriptions of experimental design in extension-driven agronomy research. Designing experiments to imitate grower standard conditions is a distinctive epistemic feature of experiments in agronomy. Grower standard does not aim to recreate so-called natural conditions. In plant biology, e.g., laboratory conditions

[2] In the United States, extension work is carried out by employees of the Cooperative Extension Service (CES), both by *county extension agents*, who manage activities for a county, and by *extension specialists*, who are academic researchers. CES is an 18,000-person agency run by the U.S. Department of Agriculture.

usually imitate native settings for plant development, without an intention that the results of the experiment will be used to change that native setting. In contrast, extension-driven agronomy experiments aim to improve grower conditions. Experimental conditions are set up as a suboptimal baseline from which to improve production, rather than as a neutral background in which scientific phenomena occur. This difference suggests a different relationship between experiment and world than in natural sciences.

Some aspects of setting grower standard are analogous to fixing variables in experimental control groups. For instance, one goal of a recent plant-pathology experiment on grape powdery mildew (*Erysiphe necator*) was to determine the efficacy of a new strategy for fungicide application in which fungicide was applied after powdery mildew spores were detected by molecular assay, rather than according to growth benchmarks or calendar alone (Thiessen, 2016). In this experiment, the authors derive their results by comparing their protocol to the standard application procedure used by vignerons for treating grape powdery mildew. The standard application is described as a control plot and contrasted with the active “detection plot”:

> Control plot fungicides were initiated at 6 inches of growth or when a risk model indicated a high risk for spore release, and detection treatment plot fungicide applications were withheld until inoculum was detected or bloom had occurred [.] Subsequent applications of fungicides followed manufacturer recommendations for reapplication depending on chemistry. [...] After a fungicide programme was initiated, additional applications in both the control and detection plots were made using the grower's standard fungicide programme. (Ibid., p. 243)

The control plots appear to be treated with grower standard protocols as evidenced by the author's description of their own experimental results, stating: "no significant differences in berry or leaf incidence between plots with fungicides initiated at detection or grower standard practice plots." (Ibid., p. 238)

We make three further observations on this case in order to thicken our description of the grower standard concept. First, the notion of a control is used in at least two distinct ways in agronomy. The first is in the way exemplified above, where grower standard practices are taken to be a contrast class for experimental interventions. The second way is to define a control as an experimental plot that receives no or very few interventions. For instance, in the experiment above, Instead of treating the control plots according to grower-standard fungicide programs, the researchers could have generated control plots with no fungicide program. Setting a no-fungicide control for that particular experiment would not have been particularly informative, since the goal of the research was to test a proposed improvement upon current standard fungicide practices.

Second, the concept of a grower standard functions in this case in ways beyond merely setting a control group for the experiment. These functions are more difficult to categorize if what we are relying on is the existing philosophical language for experiment design. In the grape powdery mildew experiment, it is evident that the notion of a grower standard guides further experimental design considerations. The experiment tests when to initiate a fungicide program, but once initiated, grower standard specifies when and how future treatments will be applied. This is somewhat similar to the role played by background conditions.

However, the protocols that see the agronomic experiment through are often carried out by growers themselves. Agronomy experiments are typically carried out on either commercial or research farms. They are designed and implemented by researchers, and are maintained by farm staff whose backgrounds are in agricultural practice rather than agricultural science. These growers are active agents in the maintenance of grower standard practices, and their practical knowledge can inform the design of agronomy experiments.

Third, the grape powdery mildew experiment demonstrates quite vividly that grower standard is not a neutral backdrop for experimental intervention. Even though there are ways in which grower standard sets background conditions for the experiment, the whole aim of the experiment is to improve upon current grower standard practices for treating powdery mildew in Oregon grapes. In this way, grower standard is conceptualized as a suboptimal baseline upon which to build improvements.

This function of grower standard is not easily recognizable in common accounts of the epistemology of experiment. We believe this is due to the difference in aims between pure and applied scientific experimentation. In pure-science experimentation, central goals of experiments are to observe, measure, detect, understand, and control natural phenomena. For instance, in the experiment to test the effects of temperature and humidity on the proliferation of grape powdery mildew, Delp (1945) concludes: “temperature is the primary factor limiting the development of vine mildew” in the regions studied during the experiment. The results of Delp’s experiment might be (and indeed, were) taken up by agronomic experimenters or by growers in

later efforts to improve growing conditions, but Delp's experiment was not framed around the investigation or improvement of grower standard.

While we do not wish to draw a hard division between pure and applied experimentation, we contend that when grower standard functions in an experiment in this baseline-setting way we describe, it does so in virtue of the applied aims of an experiment.

2.2 Grower Standard as Novel Epistemic Object

We have shown that grower standard is a complex and multi-functional concept within agronomy. It plays some familiar and some novel roles within the design and interpretation of agronomic experiments. The aim-setting and baseline-setting functions of grower standard distinguish it from both experimental controls and background conditions. We take this as evidence that grower standard is a novel epistemic object within the epistemology of experiment, that is, one that does not fit neatly into existing accounts of the phenomena and practices that comprise scientific experimentation or the epistemology of science more broadly. Grower standard is not a model, theory, instrument, type of evidence, or form of measurement. It also does not fit into the newer categories of epistemic objects suggested in recent accounts of the philosophy of scientific practice, such as Ankeny and Leonelli's repertoires (2016) or Currie's surrogate experiments and inference tools (2018).

Our analysis of grower standard shows that it is not only a novel epistemic object, but a novel *type* of epistemic object within the epistemology of experiment. For present purposes we resist the urge to name and characterize the broader category of epistemic object into which grower standard falls. However, we believe some generalizations can nonetheless be made about the

sort of epistemic object that grower standard is by further investigating relations between the functions of grower standard and the network of scientific and extra-scientific influences that interact to produce grower standard. In the next section, we discuss the relationship between grower standard and agricultural extension work.

3. Extension Work and the Epistemic Objects of Agronomy

Above, we showed that grower standard provides a non-neutral set of background conditions for experiment, that it plays a role in setting the aims and methods of experiment, and that it is not a fixed standard but rather a suboptimal baseline to be improved upon through the results of experimental intervention. In this section, we show that the existence of grower standard as an epistemic object is inextricable from consideration of how it is used by different epistemic communities as a locus for interdisciplinary exchange. First, we show that the relationship between agronomy and agricultural extension work shapes the methods for knowledge production in agronomy. Then we extend existing accounts of interdisciplinarity in the philosophy of science to lay the foundations of a framework for understanding the knowledge coproduction that occurs through agricultural extension work.

3.1 Coproducing Knowledge Through Agricultural Extension Work

Agronomy and agricultural extension work are interconnected by important contingencies of history. In the U.S., agronomy research was integrated into the mission of a group of public universities designated as the Land-Grant Institutions (LGIs). One component of the land-grant

system was to provide people an education that including agriculture, practical mechanic competencies as well as liberal arts and classics. The Hatch Act of 1887 created the agricultural experiment station program and the later 1914 the Smith-Lever Act formally associated extension work with the LGIs when it established the Cooperative Extension Service (See Footnote 2) to disseminate findings obtained from the experiment station's experiments.

The in-practice union of research and extension work means that while grower standard is *used* in agronomy experiment, it is *defined* and *known* through extension work. Growers know what grower standard practices are in practice, in their fields and with their soil. Their tacit knowledge may be shared when they show extension workers how they make decisions about when to fertilize, spray, harvest or till. Likewise, extension specialists can identify aspects of production, such as the importance of knowing that farmers will not take seriously the results of tobacco experiments planted in Kentucky after June 20, or understanding the economic impact on farmers if late-season insect control for cotton in Missouri is suspended too soon for a grower. This is one significant way in which extension work influences the epistemic objects of agronomy.

Epistemic objects like grower standard may be understood as a type of agricultural tool. Through extension work, agricultural tools can be shared, borrowed, invented, and innovated within local family farming communities, and in collaborations with research from multiple university extension centers. As with grower standard, farmers and researchers are often co-producers of these agricultural tools. These tools shape choices that farmers make about their farm and crops. For instance, given access to a mechanical seed corn harvester, a farmer might choose field corn whose ears grow at the same height facilitating more efficient picking. If

a farmer has been no-tilling her operation, she might choose to plant a cover crop of ryegrass to build up the health of the soil, especially if she has fragipan soil (Vollstedt, 2020). Tool-driven knowledge of these techniques also shapes the type of extension research that is applied to crop production, as well as affecting decisions about which experiments on test fields are performed.

We contend that agricultural extension work plays an essential role in defining a new set of epistemic categories that are essential to the practice of agricultural science. Hearkening to contemporary work on social epistemology in the sciences, we call this process the *coproduction of knowledge in agricultural science*. Importantly, the epistemic objects produced through this process are of use to both individual researchers and farmers as well as to wider populations. Further, these epistemic objects impact all of us by affecting decisions about how our food, fuel, and fiber is made.

Focusing on the knowledge-coproduction relationship between extension researchers and farmers allows us to shine a light on a central method of knowledge growth in agricultural science. We contend that this method can only be understood within the realities of extension's institutional and demographic history. As such, our nascent epistemology of agricultural extension complements current philosophical work on the contingent and value-laden epistemologies of other scientific practices. Because extension is also a formalized federal institution, we also see a particularly strong connection with current work that investigates the interplay between political and institutional pressures in shaping scientific research (e.g. Brown 2013, 2013b; Douglas, 2009; Kellert, Longino, and Waters, 2006; Kitcher, 2003, 2011; London and Zollman, 2010; Zollman, 2007).

Agricultural extension work is fundamentally an exchange of ideas between extension professionals and the communities they serve. Our analysis of the concept of grower standard shows that this exchange shapes the epistemic categories of agricultural science in an applied and interactive way. Agricultural extension has played a unique role in shaping rural and agrarian attitudes toward science. These attitudes are complex and varied, insofar as scientific innovation has greatly increased agricultural productivity, but also changed the farmer's relationship to technology, business, and state interests over the past century. Through technological innovation, it has also contributed to a diminishing agricultural. This is fertile soil for new philosophical analysis of the relationships between science, agriculture, and society.

3.2 Knowledge Coproduction in Agricultural Extension Transcends Interdisciplinary Exchange

Extension originates important epistemic objects of agricultural science, such as grower standard. But extension work is not just limited to the exchange of research and applied scientific knowledge from researcher to farmer and farmer to researcher. Extension work maps a space of communication where knowledge grows: it is the epistemic locus where a specific and impactful variety of knowledge coproduction among diverse stakeholders takes place. A robust characterization of the epistemic objects generated in extension work thus requires a deeper understanding of the standpoints of these different stakeholders, their interests, and their interactions.

It would be impractical to generate a complete taxonomy of stakeholders in extension work and agricultural science, but it is worth mentioning some common entries to illustrate the diversity of standpoints influencing the generation of epistemic objects like grower standard. We have discussed extension specialists and farmers at length already, and we have shown how farmers' interests shape grower standards. Analogous stories can be told about the interests of farmers' suppliers and consumers, as well as about institutional and funding pressures on the research programs of extension specialists. Additionally, extension work is also performed by county extension agents, whose professional obligations to research differ significantly from extension specialists, and whose training and interests likewise differ. These are all stakeholders in the shaping of epistemic objects in agricultural science.

Often, the ability to form research questions and pursue research depends on the epistemic aims and values of stakeholders within a particular agricultural environment (Bammer et al., 2013, 29-54; O'Rourke, Crowley and Gonnerman, 2016, 62-64). When philosophers have previously studied the production of epistemic objects through the collaboration among diverse stakeholders, they have primarily done so through the study of interdisciplinarity. Foundational philosophy-of-science work on interdisciplinary exchange frames interactions between disciplinarily divergent members of a scientific project as an economic exchange, specifically a "trading zone." (Galison, 1997, 1999) The metaphor is extended into linguistics by arguing that just as trading communities with different languages developed pidgin vocabularies to exchanges goods, so do scientists in different disciplines generate limited common vocabularies for the exchange of ideas, based in interactional expertise (Collins et al., 2007).

Extension work constitutes and is constituted by an interdisciplinary exchange insofar as it is knowledge that is articulated within a framework built from interactions and in-practice

experience that both shapes and is shaped by future interactions. However, extension work also seems to outstrip the notion of interdisciplinary research, due to the diversity of interests and backgrounds across stakeholders. Unlike other loci of interdisciplinary discussion, the boundary that is crossed is not just disciplinary. In extension's attempt to understand the goals and purposes of another on their own terms, what is required is more than an understanding of the position of the farm, choice of crops, and agricultural goals.

Within extension work, knowledge is always understood with reference to a particular context and in light of the actions of a number of epistemic agents. The circumscription of an epistemic object relies on how farmers use standards and tools, how these are developed in industry, the purposes for which they are used, and how each of these characteristics are informed by research within agronomy. Their use shapes diverse perceptions (within industry, university, farmer, and among consumers) and may vary depending on the crops (e.g. cotton, maize, wheat); the relationship between farmer, farm, biotech industry, society, and the environment; the interpretation of languages relied upon by farmers and scientists; and how research, technologies, and applications affect perceptions about "nature" and "cultivation." That is, an epistemic object in extension relies on a number of positionalities within academic research knowledge, applied scientific knowledges, technological knowledge, and local ecological knowledges.

Further, within extension, interactions are not limited to agent–agent interactions but include agent–object interactions as well. Knowledge coproducing interactions within extension work include researcher–farmer; farmer–veterinarian–livestock; agronomist–agrotech–banker; farmer–land; farmer–cotton baler–farm financial officer; agronomist–agricultural science research standards–university interactions; and many more. These interactions vary depending

on the crop, pest, and consumer. For instance, cotton production requires substantial up-front costs (e.g. pickers, balers), but may require less irrigation than maize. Maize may require extra irrigation around the time of tasseling. Farmers planting maize may also consider whether they will sell their crops for ethanol production or food production considering the position of the consumer and other local and global markets. In these discussions, both farmers and extension researchers are beneficiaries of the knowledge that they coproduce.

While some philosophers and historians of science have accounted for the clustering of cross-disciplinary knowledge creation around instrumentation (e.g. Mody, 2011), few have developed an account that encompasses agent–agent interdisciplinary exchange, tacit-knowledge exchanges, and what is commonly called “instrumental knowledge.” Because of the diversity of expertises and interests involved in knowledge coproduction in extension work, any epistemology of extension must incorporate all these sources of knowledge-growth interactions under a shared umbrella. This sets the knowledge-making activities of extension work apart from other sorts of knowledge-making practices in the natural sciences, and the epistemic objects created by this means are likewise distinct. Inherently defined by the ineliminable role of extension work, agronomy regularly generates epistemic objects of this experimental and interactive sort.

4. Conclusions

Knowledge coproduction in extension work and agronomy is not the result of simply applying universal rules for deriving knowledge from facts. Instead, it is the result of critical intersubjective modes of investigation between farmer and extension worker, and between farm, academy, and society. In order to illustrate what knowledge coproduction looks like within extension work, we

introduced the concept of the grower standard as an example of a coproduced epistemic object. The purpose of this was to show how knowledge is obtained through the activities of extension and communication between different stakeholders (e.g. researchers, farmer, industry, state). We showed how this form of knowledge coproduction was dependent upon these reciprocal channels of communication, and also how it transcends familiar transactional accounts of interdisciplinary research.

Although we have argued that the sorts of epistemic objects that arise from extension work are different from those arising in other disciplines, we also see strong connections between our work and other contemporary discussions in philosophy of science. In addition to literatures on interdisciplinarity and values in science, our account of grower standard as an epistemic object—as a tool that shapes and is shaped by the knowledge-making practices among a host of stakeholders—has roots in a number of different philosophical accounts of knowledge creation, including integrated history and philosophy of science, technosocial philosophy, and experimental and perspectival approaches to realism.

These authors provide motivation for our work by taking seriously the study of the interaction between humans, machines, and tools. In their views, and in ours, these interactions are the remit of a more widely extended approach to the study of philosophy of science that not only recognizes the social aspects of scientific knowledge production but sees them as ineliminable to knowledge and its growth. This approach informs the kinds of knowledge coproduction that take place within extension. In future work we hope to both jointly and individually pursue the relation between our views and these influences.

In particular, one of us will develop these foundations into a study of the normative constraints imposed on knowledge coproduction by the interests of the diverse stakeholders in extension

work. This will focus on work on the intersection of history of science, science and technology studies, and philosophy of science. Meanwhile, the other aims to compare the particularities of knowledge coproduction in extension work to knowledge coproduction in other applied sciences. As an applied science that has been historically coupled to institutional channels for communication with lay communities, the broader structure of knowledge construction in agricultural science is unlikely simply to fall in step with the structure of knowledge construction in the natural and social sciences.

We both think that the aim-setting and baseline-setting functions of grower standard also illustrate how deeply the applied aims of an experiment can be integrated with the methods of the experiment. Now-outdated views about the value-free ideal of science would suggest that this degree of integration makes for bad science, in that the data produced by the experiment are inextricable from the epistemic object of the grower standard. In future work, we will show that this degree of integration is instead an asset to agronomic experiments.

In this paper, we have provided a proof-of-concept sketch of what an epistemology of agricultural extension work might look like through our analysis of (a) grower standard as epistemic object and (b) stakeholder-driven coproduction of agronomical knowledge. We argued that agricultural science is the result of historical, social, interactive, and highly contingent agricultural practices and how the epistemic objects it produces are inextricable from those contingent histories. The example of grower standard was meant to elucidate how considerations of value are constitutive of an epistemology of experiment in agricultural science. We do not see agricultural science as an outlier, but as an archetypical instance of value-laden epistemologies in applied sciences. As such, the purpose of our paper was to prepare the

ground for future work exploring this new form of value-ladenness in the methodology of agricultural science more generally.

References

Ankeny, Rachel, and Sabina Leonelli. “Repertoires: A Post-Kuhnian Perspective on Scientific Change and Collaborative Research.” *Studies in History and Philosophy of Science*. 60 (2016): 18-28.

Bammer, Gabriele, Simon Bronitt, L. David Brown, Marcel Bursztyn, Maria Beatriz Maury, Lawrence Cram, Ian Elsum, Holly J. Falk-Krzesinski, Fasihuddin, Howard Gadlin, L. Michelle Bennett, Budi Haryanto, Julie Thompson Klein, Ted Lefroy, Catherine Lyall, M. Duane Nellis, Linda Neuhauser, Deborah O’Connell, Damien Farine, Michael O’Connor, Michael Dunlop, Michael O’Rourke, Christian Pohl, Merritt Polk, Alison Ritter, Alice Roughley, Michael Smithson, Daniel Walker, Michael Wesley, and Glenn Withers. “Front Matter.” In *Disciplining Interdisciplinarity: Integration and Implementation Sciences for Researching Complex Real-World Problems*, I-Iv. ANU Press, (2013): 29-54.

Brown, Matthew. “Values in Science Beyond Underdetermination and Inductive Risk.” *Philosophy of Science* 80, no. 5 (2013): 829-839.

———. “The Source and Status of Values for Socially Responsible Science.” *Philosophical Studies* 163, no. 1 (2013): 67-76.

Collins, Harry, Robert Evans, and Mike Gorman. “Trading Zones and Interactional Expertise.” *Studies in History and Philosophy of Science* 38, no. 4 (2007): 657-666.

Currie, Adrian. *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences*. MIT Press, 2018.

Delp, Charles. “Effect of Temperature and Humidity on the Grape Powdery Mildew Fungus.” *Phytopathology* 44, no. 11 (1954): 615-626.

Douglas, Heather. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, 2009.

Galison, Peter. *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press, 1997.

———. “Trading Zone: Coordinating Action and Belief.” *The Science Studies Reader* 13 (1999): 137-160.

Kellert, Stephen, Helen Longino, and Kenneth Waters, eds. *Scientific Pluralism*. U of Minnesota Press, 2006.

Kitcher, Philip. *Science, Truth, and Democracy*. Oxford University Press, 2003.

———. “Science in a Democratic Society.” In *Scientific Realism and Democratic Society*, pp. 95-112. Brill Rodopi, 2011.

London, Alex, and Kevin Zollman. “Research at the Auction Block: Problems for the Fair Benefits Approach to International Research.” *Hastings Center Report* 40, no. 4 (2010): 34-45.

Mody, Cyrus. *Instrumental Community: Probe Microscopy and the Path to Nanotechnology*. MIT Press, 2011.

O'Rourke, Michael, Stephen Crowley, and Chad Gonnerman. “On the Nature of Cross-Disciplinary Integration: A Philosophical Framework.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 56 (2016): 62-70.

Thiessen, Lindsey, et al. “Development of a Grower-Conducted Inoculum Detection Assay for Management of Grape Powdery Mildew.” *Plant Pathology* 65, no. 2 (2016): 238-249.

Thompson, Paul B. *The Spirit of the Soil.* Routledge, 2017.

Vollstedt, Megan. “Protecting and Improving Soil with No-Till and Cover Crops.” *Successful Farming* (2020).

Zollman, Kevin. “The Communication Structure of Epistemic Communities.” *Philosophy of Science* 74, no. 5 (2007): 574-587.

Variable Definition and Independent Components

Lorenzo Casini[*], Alessio Moneta[†], and Marco Capasso[‡]

Abstract

In the causal modelling literature, it is well known that "ill-defined" variables may give rise to "ambiguous manipulations" (Spirtes and Scheines, 2004). Here, we illustrate how ill-defined variables may also induce mistakes in causal inference when standard causal search methods are applied (Spirtes et al., 2000; Pearl, 2009). To address the problem, we introduce a representation framework, which exploits an *independent component representation* of the data, and demonstrate its potential for detecting ill-defined variables and avoiding mistaken causal inferences.

1 The problem of variable definition

Some choices of variables may lead to less informative, or even false, causal claims. This problem was pointed out by, among others, Spirtes and Scheines (2004), Eberhardt (2016), and Woodward (2016). Here is a classic example by Spirtes and Scheines (2004). Consider the following hypothetical data generating process (Figure 1). Total cholesterol (TC) is a deterministic function (e.g., the sum) of two variables, viz. low-density lipoproteins (LDL) and high-density lipoproteins (HDL), respectively known as "bad" and "good" cholesterol. The two cholesterols, in fact, have different causal roles: LDL causes heart disease (HD), while HDL prevents it. Moreover, assume that HDL and LDL cause, respectively, a disease called "disease 1" ($D1$) and a disease called "disease 2" ($D2$). Spirtes and Scheins point out that, if only TC, but neither HDL nor LDL is observed, a manipulation of TC with respect to HD is "ambiguous", because it leaves underdetermined the values of TC's underlying determinants, such that the effect on HD is unpredictable.

[*]Department of Philosophy, University of Geneva, Switzerland. Email: *lorenzo.casini@unige.ch*

[†]Institute of Economics, Scuola Superiore Sant'Anna, Pisa, Italy. Email: *a.moneta@santannapisa.it*

[‡]Nordic Institute for Studies in Innovation, Research and Education, Oslo, Norway. Email: *marco.capasso@gmail.com*

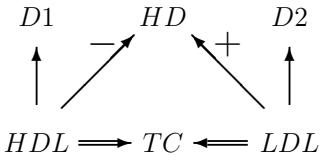


Figure 1: A structure where the manipulation on TC with respect to HD is "ambiguous".

More generally, in applied causal inference, often the variables under study are, like TC, functions of other variables with heterogeneous causal roles. For example, in macroeconomics a researcher deals with aggregate variables such as gross domestic product, foreign sales, total imports, etc., which are sums or averages of other variables, whose individual causal roles may be multifarious and opaque to the researcher. Often, the researcher is unable to observe the underlying micro-behaviours simply because statistical agencies provide aggregate data, but do not reveal information on the single units. In other cases, collecting micro data may be too complex or costly. Treating aggregate variables as if they had a homogeneous causal role, however, may lead to less informative or false causal claims, as shown by the TC example. We shall refer to an aggregate variable incurring such problems as *ill-defined*. Notice, thus, that whether a variable is ill-defined is relative to a variable set. That is, it may be ill-defined in a set but well-defined in another.

The problem of variable definition is often underestimated by the wider public. For instance, not sufficient attention has been paid to its consequences for causal inference by constraint-based discovery methods (Spirtes et al., 2000; Pearl, 2009). We shall return to this point in the next section, by showing how the presence of TC in a variable set may lead to wrong causal inferences. To address the problem, we introduce a representation framework—the "independent component" representation—for modelling structures containing two kinds of dependencies, namely traditional causal dependencies between well-defined variables, and dependencies between ill-defined variables and their determinants (see, e.g., Figure 1). Next, we demonstrate the potential of this framework for identifying ill-defined variables and reducing the risk of mistaken causal inferences.

2 Causal search with ill-defined variables

The last decades have witnessed the development and popularization of constraint-based discovery methods for causal inference (Spirtes et al., 2000; Pearl, 2009).

In this framework, a causal structure is represented as a triple $\langle \mathbf{V}, \mathcal{E}, \Pr \rangle$, where $\langle \mathbf{V}, \mathcal{E} \rangle$ is a directed acyclic graph (DAG) consisting of a set $\mathbf{V}$ of variables and a set $\mathcal{E}$ of edges among them, and $\Pr$ is the probability distribution over $\mathbf{V}$ associated to the DAG. $\Pr$ is assumed to comply with the Causal Markov Condition (CMC) and, typically, the Causal Faithfulness Condition (CFC). CMC says that

(CMC) For any $V_i \in \mathbf{V} = \{V_1, \ldots, V_n\}$, $V_i \perp\!\!\!\perp \mathbf{Non_i} | \mathbf{Par_i}$,

where $\mathbf{Par_i}$ denotes the set of parents (direct causes) of V_i, and $\mathbf{Non_i}$ denotes the set of non-descendants (non-effects) of V_i. In words, each variable is probabilistically independent of its non-effects, conditional on its direct causes. CMC presupposes that for every pair of variables in $\mathbf{V}$, every common direct cause of the pair is in $\mathbf{V}$ or has the same value for all units in the population (causal sufficiency). CFC says:

(CFC) $\langle \mathbf{V}, \mathcal{E}, \Pr \rangle$ is such that every conditional independence relation true in $\Pr$ is entailed by CMC applied to the true DAG $\langle \mathbf{V}, \mathcal{E} \rangle$.

CFC ensures that there is no causal dependence without probabilistic dependence, that is, all probabilistic independencies in the DAG correspond to causal independencies.

Based on these assumptions, constraint-based discovery methods are designed to recover the causal structure from data, by identifying conditional independencies among variables and then causally connecting variables not found to be independent. We shall now consider examples of simple data generating processes including one ill-defined variable, TC, and show how using constraint-based methods based on conditional independencies—whilst ignoring that TC is ill-defined—may lead to mistakes. To anticipate, such mistakes involve apparent violations of CMC or CFC, which the search methods presuppose. Notice, however, that our interest here is not in providing novel counterexamples to CMC and CFC. These violations, in fact, could be avoided by choosing a "more suitable" variable set for causal inference—in this case, one featuring HDL and LDL instead of TC. And indeed, a formulation of CMC requiring that variables be independent of their non-effects conditional on their *well-defined* direct causes would not incur any violation. In this paper, however, we do not want to presuppose what counts as an ill-defined variable or a suitable variable set. Our goal is to avoid mistaken causal inferences *in virtue of detecting ill-defined variables*.

Suppose that, in $\mathbf{V} = \{X, Y, Z\}$, Y is the non-deterministic cause of both X and Z, viz. the true structure is $X \longleftarrow Y \longrightarrow Z$. If all variables are well-defined,

one can infer some properties of the causal structure by testing conditional independencies and applying a constraint-based discovery method. In particular, the independence $X \perp\!\!\!\perp Z \,|\, Y$ and CFC allow one to exclude $X \longrightarrow Y \longleftarrow Z$ from the set of possible structures. Now, let the set of observed variables be $\mathbf{V}' = \{TC, D1, D2\}$. That is, suppose again that one does not observe or measure LDL and HDL, but only TC. In this case, too, the true structure is not a collider. Assuming that the dependencies over $\mathbf{V}'$ are causally interpretable, the most plausible structure—the one we wish to rationalize in this paper—would be a common cause, viz. $D1 \longleftarrow TC \longrightarrow D2$. However, since HDL and LDL are independent, $LDL \perp\!\!\!\perp HDL$, it follows that $D1$ and $D2$ are independent, too, viz. $D1 \perp\!\!\!\perp D2$. If the true structure is a common cause, this contradicts CFC, which would entail a dependence between the effects of the common cause. Moreover, being $D1$ and $D2$ dependent on (respectively) LDL and HDL, $D1$ and $D2$ become dependent upon conditioning on TC, viz. $D1 \not\perp\!\!\!\perp D2 \,|\, TC$. For example, suppose one knows that one patient's total cholesterol has increased. Then, knowing that disease 1 is absent gives one relevant information to predict that disease 2 is present. If the true structure is a common cause, this conditional dependence would violate CMC, which would entail the independence of $D1$ and $D2$ given their common cause. Based on $D1 \perp\!\!\!\perp D2$ and $D1 \not\perp\!\!\!\perp D2 \,|\, TC$, as well as $TC \not\perp\!\!\!\perp D1$ and $TC \not\perp\!\!\!\perp D2$, a constraint-based algorithm (e.g., PC, FCI; Spirtes et al. 2000) will infer an unshielded collider on TC, viz. $D1 \longrightarrow TC \longleftarrow D2$. A researcher applying the algorithm without knowing that TC is the sum of HDL and LDL (which are causes of, respectively, $D1$ and $D2$) will thus infer the wrong structure. The reason, ultimately, is that TC is ill-defined in $\mathbf{V}'$.

Similarly, assume that all variables in $\mathbf{V}$ are well-defined, but now X causes Y, and Y causes Z, viz. the true structure is $X \longrightarrow Y \longrightarrow Z$. Under CMC, it holds $Z \perp\!\!\!\perp X \,|\, Y$, and under CFC, it holds $X \not\perp\!\!\!\perp Z$. Now, consider the set of observed variables $\mathbf{V}'' = \{Da, TC, D1\}$, where Da (not represented in Figure 1), denoting dairies, is a cause of LDL but not of HDL. Again, suppose that one observes TC but neither HDL nor LDL. Here, too, the true structure is not a collider. The most plausible causal interpretation of the dependencies over $\mathbf{V}''$ is a directed path, viz. $Da \longrightarrow TC \longrightarrow D1$. However, since Da is a cause of LDL, which is independent of the cause HDL of $D1$, it holds $Da \perp\!\!\!\perp D1$, which violates CFC. Moreover, it holds $Da \not\perp\!\!\!\perp D1 \,|\, TC$, which violates CMC. From this, one may again wrongly infer a collider on TC, viz. $Da \longrightarrow TC \longleftarrow D1$. Ultimately, the reason is that TC ill-defined in $\mathbf{V}''$.

These simple examples show how conditional independencies are sensitive to

the presence of ill-defined variables in fork and chain structures[1] but ill-defined variables are undetectable from conditional independencies only. This may lead to mistaken inferences (viz. the inference of colliders) if one unreflectively applies constraint-based algorithms.

3 A novel representation framework

We now introduce a series of definitions, which will allow us to precisely define the notion of ill-defined variable. First, we introduce a class of data generating mechanisms inducing the problem of ill-defined variables. We call them "augmented" structural causal models, by which we extend the traditional notion of structural causal models (Pearl, 2009; Peters et al., 2017) to structures including deterministic assignments.

Augmented structural causal model An *augmented structural causal model* $\mathfrak{C} := (\mathbf{A}_W, \mathbf{A}_I, \Pr)$ consists of a collection $\mathbf{A}_W$ of m assignments, a collection $\mathbf{A}_I$ of k assignments, and a probability distribution $\Pr$ such that:

(i) the collection of $\mathbf{A}_W$ consists of assignments

$$W_i := f_i(\mathbf{Par_i}, S_i), \qquad \text{for } i = 1, \ldots, m,$$

where $\mathbf{Par_i} \subseteq \mathbf{W} \backslash \{W_i\}$ are called the parents of W_i, and S_i are called *noises*, or *shocks*;

(ii) $\Pr$ over $\mathbf{S} = \{S_1, \ldots, S_m\}$ is such that the shocks are mutually independent, viz. $\Pr(\mathbf{S}) = \Pr(S_1) \cdot \ldots \cdot \Pr(S_m)$; hence, the S_i are also called *independent components*;

(iii) the collection of $\mathbf{A}_I$ consists of assignments

$$I_i := f_i(\mathbf{Det_i}), \qquad \text{for } i = 1, \ldots, k,$$

where $\mathbf{Det_i} \subset \mathbf{V}$ are called *determinants* of I_i.

$\mathfrak{C}$ is defined over a set of variables $\mathbf{V} = \mathbf{W} \cup \mathbf{I}$ with cardinality $n = m + k$. We associate to $\mathfrak{C}$ a graph $\mathcal{G}_V$ (see, e.g. the graph in Figure 1, where TC is the only variable with a deterministic assignment). $\mathcal{G}_V$ is obtained by creating a node for

[1] By contrast, no mistake occurs if TC is truly a collider. For instance, the inferred structure over $\mathbf{V}''' = \{Da, TC, Ol\}$, where Ol (olive oil) causes HDL but not LDL, is $Da \longrightarrow TC \longleftarrow Ol$, as it should be.

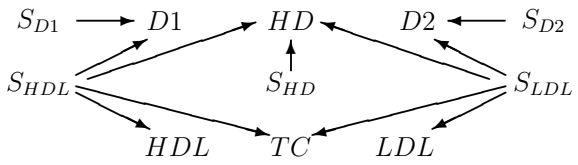


Figure 2: $\mathcal{G}_V^{IC}$ corresponding to the DAG $\mathcal{G}_V$ in Figure 1.

each element of $\mathbf{V}$, and by drawing a directed edge $\longrightarrow$ from each parent in $\mathbf{Par_i}$ (if not empty) to W_i, and a *modified* directed edge $\Longrightarrow$ from each determinant in $\mathbf{Det_i}$ to I_i. Henceforth, we restrict our attention to acyclic structures, such that $\mathcal{G}_V$ is a *modified* DAG, to cases where $\mathbf{Det_i}$ has at least two elements, and to assignments $\mathbf{A}_W$ in which the shocks are additive. For simplicity, we also assume that no pair of variables I_i, I_j in $\mathbf{I}$, $I_i \neq I_j$, are linked in $\mathcal{G}_V$ by a bidirected modified "active" (i.e., without colliders) path $I_i \Longleftarrow \cdots \Longrightarrow I_j$.

By replacing all modified directed edges $\Longrightarrow$ with standard edges $\longrightarrow$, $\mathcal{G}_V$ becomes a standard DAG, labelled $\widetilde{\mathcal{G}}_V$. By removing from $\mathcal{G}_V$ the nodes in $\mathbf{I}$ and the edges connecting $\mathbf{I}$ to $\mathbf{W}$, we obtain a subgraph of $\mathcal{G}_V$, which we denote $\mathcal{G}_W$.

Let us now introduce a particular graph associated with $\mathfrak{C}$, which we call independent component (IC) representation, or $\mathcal{G}^{IC}$. $\mathcal{G}^{IC}$ contains edges between shocks and endogenous variables but not among endogenous variables themselves. Despite this apparent limitation, the information in $\mathcal{G}^{IC}$ shall be key to the purpose of our paper. Although here we are not concerned with how $\mathcal{G}^{IC}$ is recovered, we should mention that there exist powerful statistical learning techniques, such as Independent Component Analysis (ICA) (Hyvärinen et al., 2001), which under certain assumptions (viz., non-Gaussianity) infer the dependence coefficients, and thus identify the absence of dependencies, between shocks and endogenous variables in $\mathfrak{C}$, and thereby recover the edges in $\mathcal{G}^{IC}$.

IC representation Consider $\mathfrak{C} := (\mathbf{A}_W, \mathbf{A}_I, \Pr)$, with $\mathbf{V} = \mathbf{W} \cup \mathbf{I}$, $card(\mathbf{V}) = n = m+k$. An *IC representation* of $\mathfrak{C}$ is a DAG $\mathcal{G}_V^{IC} = \langle \mathbf{V} \cup \mathbf{S}, \mathcal{E}^{IC} \rangle$ such that $\mathcal{E}^{IC}$ consists of the following edges: (i) $S_i \longrightarrow W_i$, for any $i = 1, \ldots, m$; (ii) $S_i \longrightarrow W_j$, for any $i \neq j$ such that there is a directed standard path $W_i \longrightarrow \cdots \longrightarrow W_j$ in $\mathcal{G}_V$; (iii) $S_i \longrightarrow I_h$, for any $S_i \in \mathbf{S}$ and any $I_h \in \mathbf{I}$ such that there is a directed modified path $W_i \Longrightarrow \cdots \Longrightarrow I_h$ in $\mathcal{G}_V$; (iv) $S_i \longrightarrow I_h$, for any $S_i \in \mathbf{S}$ and any $I_h \in \mathbf{I}$ such that from W_i to I_h in $\mathcal{G}_V$ there is a directed standard path followed by a directed modified path with the same orientation, $W_i \longrightarrow \cdots \Longrightarrow I_h$.

Let us illustrate this definition relative to Figure 2, where $\mathbf{W} = \{HDL, LDL, D1, D2, HD\}$ and $\mathbf{I} = \{TC\}$. (i) There is a shock for each variable in $\mathbf{W}$. Some

shocks (e.g., S_{D1}) only hit one variable ($D1$). Other are common to multiple variables. (ii) For any variable (e.g., HDL), its shock (S_{HDL}) also hits all of its descendants, if any ($D1, HD$). (iii) Any shock to a determinant of a variable I_i in $\mathbf{I}$ (e.g., S_{HDL}) also hits I_i (TC). (iv) If $\mathbf{V}$ contained a cause of a determinant of I_i (e.g., dairies, Da, which causes LDL), its shock (S_{Da}) would also hit I_i (TC).

One may also define $\mathcal{G}^{IC}$ relative to any subset $\mathbf{O}$ of variables in $\mathbf{V}$, namely $\mathcal{G}^{IC}_O = \langle \mathbf{O} \cup \mathbf{S}_O, \mathcal{E}^{IC}_O \rangle$. $\mathbf{S}_O$ is obtained by removing from $\mathbf{S}$ those shocks, which $\mathfrak{C}$ assigns to variables in $\mathbf{W}$ that are not in $\mathbf{O}$, and by adding those shocks, which $\mathfrak{C}$ assigns to variables in $\mathbf{W}$ that are determinants of variables in $\mathbf{I} \cap \mathbf{O}$. $\mathcal{E}^{IC}_O$ is obtained by removing from $\mathcal{E}^{IC}$ all of those edges, whose tails are not in $\mathbf{S}_O$. For any variable set $\mathbf{O}$, we call "idiosyncratic" a shock to a variable X in $\mathcal{G}^{IC}_{\mathbf{O}}$ that is a parent of X and of no other variable. We may now define *ill-* and *well*-defined variables:

Ill- and well-defined variables Let $\mathfrak{C}$ over $\mathbf{V} = \mathbf{W} \cup \mathbf{I}$ contain the assignment $I := f(\mathbf{Det}_I)$, $card(\mathbf{Det}_I) \geq 2$. Let $\mathbf{Des}_I$ denote the set of all descendants of determinants of I in $\mathcal{G}_V$.[2] Assume $I \in \mathbf{O} \subseteq \mathbf{V}$. Then, I is *ill-defined* in $\mathbf{O}$ if and only if, for some $Des_j \in \mathbf{Des}_I$, there exists a variable Y such that (i) $Y \in \mathbf{O}$, (ii) $Y \neq I$, (iii) Y belongs to a (possibly empty) active path from Det_i to Des_j in $\mathcal{G}_V$ (viz. $Det_i \longrightarrow \cdots \longrightarrow Des_j$ or $Det_i \longrightarrow \cdots \Longrightarrow Des_j$), and (iv) $\mathcal{G}^{IC}_{\{I,Y\}}$ contains no shock S_Y common to I, Y, for which $S_Y \perp\!\!\!\perp Y | I$ in $\mathfrak{C}$. Any variable in $\mathbf{O}$ that is not ill-defined in $\mathbf{O}$ is *well-defined* in $\mathbf{O}$.

For instance, TC is well-defined in $\{Da, TC\}$ because Da is neither a determinant of TC nor a descendant of a determinant of TC, and vice versa. By contrast, TC is ill-defined in $\{HDL, TC\}$ because HDL is a determinant of TC, and $S_{HDL} \not\perp\!\!\!\perp HDL | TC$; also, TC is ill-defined in $\{TC, D1\}$ and $\{TC, HD\}$ because $D1$ and HD are effects of determinants of TC, and (respectively) $S_{D1} \not\perp\!\!\!\perp D1 | TC$ and $S_{HD} \not\perp\!\!\!\perp HD | TC$. More generally, a variable I is ill-defined in $\mathbf{O}$ if and only if $\mathbf{O}$ also contains a variable Y among I's determinants or their descendants, and $\mathcal{G}^{IC}_{\{I,Y\}}$ contains no shock S_Y on I, Y, such that I screens off S_Y from Y in $\mathfrak{C}$. This lack of screening off intuitively captures the idea that a manipulation of I with respect to Y is ambiguous. In turn, to explain the lack of screening off, we need the following Proposition (proof in Appendix):

Proposition 1 Let $\mathfrak{C}$ over $\mathbf{V} = \mathbf{W} \cup \mathbf{I}$ contain the assignment $I := f(\mathbf{Det}_I)$, $card(\mathbf{Det}_1) \geq 2$. Assume CMC and CFC in $\mathcal{G}_W$. Then, for any Det_i, Des_i, Anc_i,

[2] Notice that $\mathbf{Det}_I \subseteq \mathbf{Des}_I$ by definition of "descendant".

where Des_i is a descendant of Det_i, and Anc_i is an ancestor of Det_i, it holds $Anc_i \not\perp\!\!\!\perp Des_i | I$, except for a parameter set Θ (characterising the assignments in $\mathfrak{C}$) that violates CFC in $\widetilde{\mathcal{G}}_V$.[3]

We can also define a graph $\mathcal{G}_O = \langle \mathbf{O}, \mathcal{E}_O \rangle$ representing the structure over $\mathbf{O}$, where $\mathcal{E}_O$ consists of the following edges. First, $\mathcal{G}_O$ has a modified edge $X \Longrightarrow Y$ if and only if there is a directed path $X \Longrightarrow \cdots \Longrightarrow Y$ in $\mathcal{G}_V$, and no variable between X and Y is in $\mathbf{O}$. Next, let the tail $\diamondsuit$ of the arrow $X \diamondsuit\!\!\longrightarrow Y$ indicate that X is ill-defined in $\{X, Y\}$. Then, $\mathcal{G}_O$ has an edge $X \diamondsuit\!\!\longrightarrow Y$ for any $\langle X, Y, Z \rangle$ for which $X, Y \in \mathbf{O}$, $Z \in \mathbf{V}$, $Z \notin \mathbf{O}$, and $\mathcal{G}_V$ features a path $X \Longleftarrow Z \longrightarrow Y$, *unless* $\mathcal{G}_O^{IC}$ has a shock S common to X, Y for which $S \perp\!\!\!\perp Y | X$ in $\mathfrak{C}$, in which case $X \longrightarrow Y$ is in $\mathcal{G}_O$. Furthermore, $\mathcal{G}_O$ has a standard edge $X \longrightarrow Y$ if $\mathcal{G}_V$ has a directed path from X to Y featuring standard edges $\longrightarrow$ and/or modified edges $\Longrightarrow$, and no variable between X and Y is in $\mathbf{O}$. Finally, $\mathcal{G}_O$ has a bidirected edge $X \longleftrightarrow Y$ if and only if $\mathcal{G}_V$ has an active path $X \longleftarrow \cdots \longleftarrow Z \longrightarrow \cdots \longrightarrow Y$ featuring standard or modified edges, and only X, Y on that path are in $\mathbf{O}$. No further edges are in $\mathcal{G}_O$.

Illustrated in relation to Figure 1, $\mathcal{G}_{\{HDL,TC,LDL\}}$ is $HDL \Longrightarrow TC \Longleftarrow LDL$, and $\mathcal{G}_{\{HDL,LDL,HD\}}$ is $HDL \longrightarrow HD \longleftarrow LDL$. The two problematic structures with ill-defined variables from §2, namely $\mathcal{G}_{\{TC,D1,D2\}}$ and $\mathcal{G}_{\{Da,TC,D1\}}$, are represented as, respectively, $D1 \longleftarrow\!\!\diamondsuit TC \diamondsuit\!\!\longrightarrow D2$ and $Da \longrightarrow TC \diamondsuit\!\!\longrightarrow D1$. Finally, let us define the notions of ill- and well-defined *causes*:

Ill- and well-defined causes For any $X, Y \in \mathbf{O}$, X is an *ill-defined cause* of Y in $\mathbf{O}$ if and only if $\mathcal{G}_{\{X,Y\}}$ contains the edge $X \diamondsuit\!\!\longrightarrow Y$. For any $X, Y \in \mathbf{O}$, X is a *well-defined cause* of Y in $\mathbf{O}$ if and only if Y is well-defined in $\{X, Y\}$, and $\mathcal{G}_{\{X,Y\}}$ contains the edge $X \longrightarrow Y$.

For instance, HDL is a well-defined cause of HD in $\{HDL, HD\}$.[4] By contrast, TC is an ill-defined cause of HD in $\{TC, HD\}$.

4 Identification

We now illustrate the applicability of our framework to detecting ill-defined variables and improving causal inference. We begin with a condition, under which

[3] Notice that we do *not* assume CFC in $\widetilde{\mathcal{G}}_V$. For such a Θ, I counts as well-defined in our framework, as the manipulation of I with respect to Des_i is not ambiguous.

[4] At the same time, HDL is not a (well-defined) cause of TC in $\{HDL, TC\}$, because TC is not well-defined in that set.

one may unambiguously identify ill-defined variables.

Proposition 2: Sufficient condition for ill-definedness Consider $\mathfrak{C}$ over $\mathbf{V}$, and $\mathbf{O} = \{X, Y, Z\} \subseteq \mathbf{V}$. Assume CMC and CFC in $\mathcal{G}_W$. Also assume (i) $X \perp\!\!\!\perp Z$, (ii) $X \not\perp\!\!\!\perp Y$, $Y \not\perp\!\!\!\perp Z$, $X \not\perp\!\!\!\perp Z|Y$, and (iii) $\mathcal{G}^{IC}_O$ has no idiosyncratic shock on Y. Then, Y is ill-defined in $\mathbf{O}$ with two determinants in $\mathbf{V}$, and $\mathcal{G}_O$ is $X \leftarrow\!\!\diamond\, Y \,\diamond\!\!\rightarrow Z$.

For instance, applied to $\mathbf{V}' = \{TC, D1, D2\}$, this condition establishes that TC is an ill-defined common cause of $D1$ and $D2$, viz. $D1 \leftarrow\!\!\diamond\, TC \,\diamond\!\!\rightarrow D2$, since $D1 \perp\!\!\!\perp D2$, $D1 \not\perp\!\!\!\perp TC$, $TC \not\perp\!\!\!\perp D2$, $D1 \not\perp\!\!\!\perp D2|TC$, and $\mathcal{G}^{IC}_{V'}$ has no idiosyncratic shock to TC. Proposition 2 is easily generalizable to cases with more than two determinants.

If one observes no effects of independent determinants of the ill-defined variable, for instance in $\mathbf{V}'' = \{Da, TC, D1\}$, the above condition is not applicable. Nonetheless, one may still reduce the ambiguity concerning ill-defined variables and partially recover the causal structure. To this end, let us assume that *d*eterminism induces *d*ependencies (DD):

(DD) For any I and any $Det_i \in \mathbf{Det}_I$ in $\mathfrak{C}$, it holds $I \not\perp\!\!\!\perp Det_i$.

In words, there are probabilistic dependencies between variables with deterministic assignments and their determinants. This assumption is only violated by cancelling paths from determinants to determined variables. Its satisfaction requires (similarly to CFC) the absence of special parameterizations. For simplicity, we also assume that $\mathbf{O}$ contains no determinants of variables in $\mathbf{O}$, such that $\mathcal{E}_O$ contains no modified edges $\Longrightarrow$.[5] Then, one may identify well-defined variables:

Proposition 3: Sufficient condition for well-definedness Consider $\mathfrak{C}$ over $\mathbf{V}$, and $\mathbf{O} \subseteq \mathbf{V}$. Assume DD. Assume CMC and CFC in $\mathcal{G}_W$. Assume that no determinant of ill-defined variables in $\mathbf{O}$ is in $\mathbf{O}$. Then, a variable X is well-defined in $\mathbf{O}$ if for any Y in $\mathbf{O}$, $X \neq Y$, one of (i)–(iv) holds: (i) $X \perp\!\!\!\perp Y$; (ii) in $\mathcal{G}^{IC}_{\{X,Y\}}$ X is not a child of an idiosyncratic shock, and X, Y are children of a common shock S, such that $S \perp\!\!\!\perp Y|X$; (iii) in $\mathcal{G}^{IC}_{\{X,Y\}}$ X is the only child of an idiosyncratic shock; (iv) in $\mathcal{G}^{IC}_{\{X,Y\}}$, X, Y are children of idiosyncratic shocks, and there is $\mathbf{Z} \subset \mathbf{O}$ such that $X \perp\!\!\!\perp Y|\mathbf{Z}$ and no $Z_i \in \mathbf{Z}$ is the child of an idiosyncratic shock in $\mathcal{G}^{IC}_{\{X,Z_i\}}$.

[5] Of course, there is no *a priori* guarantee that $\mathbf{O}$ contains no determinants. Although one could easily relax this assumption, and thereby obtain a more general result, this would require a lengthier proof. For reasons of space, here we prioritize simplicity over generality.

For instance, Da (which, to recall, causes LDL but not HDL) is well-defined in $\mathbf{V}''$, since (i) $Da \perp\!\!\!\perp D1$, and (ii) $\mathcal{G}^{IC}_{\{Da,TC\}}$ contains a shock S common to Da, TC, such that $S \perp\!\!\!\perp TC|Da$, and no idiosyncratic shock to Da, from which one may infer $Da \longrightarrow TC$. Next, one can identify putative ill-defined variables:

Proposition 4: Necessary condition for ill-definedness Consider $\mathfrak{C}$ over $\mathbf{V}$ and its associated graph $\mathcal{G}_V$. Assume DD. Assume CMC and CFC in $\mathcal{G}_W$. Let X be ill-defined in $\mathbf{O} = \{X, Y\}$ with $\mathbf{Det}_X \cap \mathbf{O} = \emptyset$, $\mathbf{O} \subseteq \mathbf{V}$. Then: (i) $X \not\perp\!\!\!\perp Y$; (ii) in $\mathcal{G}^{IC}_O$ X, Y are children of a common shock; (iii.a) in $\mathcal{G}^{IC}_O$ X is child of an idiosyncratic shock, *or* (iii.b) in $\mathcal{G}^{IC}_O$ X is not a child of an idiosyncratic shock and there is a set of shocks $\mathbf{S}$ on X such that $X \perp\!\!\!\perp Y|\mathbf{S}$.

For instance, TC and $D1$ are such that (i) $TC \not\perp\!\!\!\perp D1$. Moreover, in $\mathcal{G}^{IC}_{\{TC,D1\}}$ they are (ii) children of a common shock and (iii.a) children of idiosyncratic shocks. Therefore, TC and $D1$ qualify as putatively ill-defined. Assuming the absence of bidirected modified paths, $\mathcal{G}_{\{TC,D1\}}$ cannot be $TC \Longleftarrow \cdots \Longrightarrow D1$. Therefore, only three structures are possible, namely $TC \diamondsuit\!\!\longrightarrow D1$, $TC \longleftarrow\!\!\diamondsuit D1$, and $TC \longleftrightarrow D1$. The ambiguity may be resolved by enlarging $\mathbf{V}''$ until a sufficient set $\mathbf{Z}$ of common causes of $TC, D1$ is found that screens them off, or (given $\mathbf{Z}$) the dependence between TC and $D1$ is oriented such that one is a well-defined cause of the other, viz. $TC \longrightarrow D1$ or $TC \longleftarrow D1$, or enough effects of determinants of TC or $D1$ are observed as to remove the idiosyncratic shock on TC or $D1$, such that either $TC \diamondsuit\!\!\longrightarrow D1$ or $TC \longleftarrow\!\!\diamondsuit D1$ holds.

5 Conclusion

The problem of variable definition is known to be responsible for ambiguous manipulations. Furthermore, we showed that it can lead to mistakes in causal inference by standard constraint-based causal search methods. To address the problem, we introduced a novel representation framework suitable for structures including ill-defined variables, viz. the independent component (IC) representation. We argued that recovering the IC representation can unambiguously identify ill-defined variables, under certain assumptions, or at least exclude that certain variables are ill-defined, and consequently reduce the risk of mistaken causal inferences. Given recent advances in statistical techniques (e.g., Independent Componenent Analysis) by which one may recover the IC representation, our proposal holds great promise. Therefore, we strongly invite further research on the subject.

Appendix

Proof of Proposition 1. Assume *per absurdum* that there exist Anc_i, Des_i of Det_i, such that $Anc_i \perp\!\!\!\perp Des_i | I$ for any set of parameters Θ in $\mathfrak{C}$. This is possible only if one of (A)–(C) holds: **(A)** Det_i suffices to determine I, such that I renders Det_i irrelevant to Anc_i, Des_i. This requires *card*$(\mathbf{Det}_I) = 1$, contradicting *card*$(\mathbf{Det}_I) \geq 2$. **(B)** *card*$(\mathbf{Det}_I) \geq 2$ and for some $Det_j \in \mathbf{Det}_I$, there is no directed path $Det_j \longrightarrow \cdots \longrightarrow Des_i$. Then, Det_j would act as an exogenous noise on I, such that the edge $Det_i \Longrightarrow I$ would be observationally indistinguishable from a standard edge $Det_i \longrightarrow I$. Holding CFC in $\mathbf{W}$, and since I behaves like a child of Det_i, we would have $Anc_i \not\perp\!\!\!\perp Des_i | I$, contradicting our starting hypothesis. **(C)** *card*$(\mathbf{Det}_I) \geq 2$ and for any $Det_j \in \mathbf{Det}_I$, there is a directed path $Det_j \longrightarrow \cdots \longrightarrow Des_i$. Then, there exists a parameter set Θ such that $Anc_i \perp\!\!\!\perp Des_i | I$ and, necessarily, for any $Det_i, Det_j \in \mathbf{Det}_I$, $P_\Theta(I, Des_i | Det_i) = P_\Theta(I, Des_i | Det_j)$. For instance, assume *card*$(\mathbf{Det}_I) = 2$ and a generalized additive model such that $I = f(Det_i) + g(Det_j)$ and $D = f'(Det_i) + g'(Det_j) + S_D$. Then, $A \perp\!\!\!\perp D | I$ holds only if $f(Det_i) + f'(Det_i) = g(Det_i) + g'(Det_i)$. This point generalizes to larger cardinalities. Finally, since I is a parent of neither Anc_i nor Des_i in $\widetilde{\mathcal{G}}_V$, any parameter set Θ such that $Anc_i \perp\!\!\!\perp Des_i | I$ necessarily violates CFC in $\widetilde{\mathcal{G}}_V$. □

Proof of Proposition 2. Let $*\!\!\longrightarrow$ denote one among $\longrightarrow$, $\longleftrightarrow$, and $\diamondsuit\!\!\longrightarrow$. Assume *per absurdum* that (i)–(iii) are true but Y is well-defined. CMC and (ii) entail that $\mathcal{G}_V$ contains paths linking X, Y and Y, Z. CFC and (i) entail that $\mathcal{G}_V$ contains no path linking X, Z. Then, $\mathcal{G}_O$ contains only two edges, one connecting X, Y, and one connecting Y, Z. Among the possible structures in $\mathcal{G}_O$, $X *\!\!\longrightarrow Y \longrightarrow Z$, $X \longleftarrow Y \longleftrightarrow\!\!* Z$, $X \longleftrightarrow\!\!* Y \longrightarrow Z$, and $X \longleftarrow Y *\!\!\longrightarrow Z$ contradict (i), and $X *\!\!\longrightarrow Y \longleftrightarrow\!\!* Z$ contradicts (iii). In all other structures, viz. $X \longleftarrow\!\!\diamondsuit Y \diamondsuit\!\!\longrightarrow Z$, $X *\!\!\longrightarrow Y \diamondsuit\!\!\longrightarrow Z$, and $X \longleftarrow\!\!\diamondsuit Y \longleftrightarrow\!\!* Z$, Y is ill-defined. The latter two contradict (iii). Thus, $\mathcal{G}_O$ is $X \longleftarrow\!\!\diamondsuit Y \diamondsuit\!\!\longrightarrow Z$, and $\mathbf{Det}_Y$ has precisely two elements in $\mathbf{V}$ (one causing X and one causing Z); otherwise $\mathcal{G}_O^{IC}$ would contain an idiosyncratic shock on Y associated to its extra determinant(s), violating (iii). As a corollary, $\mathcal{G}_O^{IC}$ contains idiosyncratic shocks on X and Z. □

Proof of Proposition 3. **(i)** From the definition of ill-defined variable, for any $I \in \mathbf{V}$, $\mathcal{G}_V$ contains a directed path from some $Det_i \in \mathbf{Det}_I$ to some descendant Des_j of Det_i. Under CFC and DD, I is ill-defined only if $\mathbf{O}$ contains some Y on that path, such that $I \not\perp\!\!\!\perp Y$. Hence, if $\mathbf{O} = \{X, Y\}$ and $X \perp\!\!\!\perp Y$, then X is well-defined. **(ii)** In $\mathcal{G}^{IC}_{\{X,Y\}}$, X is ill-defined and not a child of an idiosyncratic shock only if $\mathcal{G}_V$ contains directed paths from each $Det_i \in \mathbf{Det}_X$ to Y. Then, $\mathcal{G}^{IC}_{\{X,Y\}}$

contains no shock S common to X, Y, such that $S \perp\!\!\!\perp Y|X$. Since this contradicts (ii), X cannot be ill-defined. **(iii)** If X is the only child of an idiosyncratic shock in $\mathcal{G}^{IC}_{\{X,Y\}}$, then $\mathcal{G}^{IC}_{\{X,Y\}}$ contains a shock common to X, Y. Then, X is ill-defined in $\mathcal{G}^{IC}_{\{X,Y\}}$ only if $\mathbf{O}$ contains a node $Det_i \in \mathbf{Det}_X$, which is not a child of an idiosyncratic shock. This contradicts the assumption that $\mathbf{Det}_X \cap \mathbf{O} = \emptyset$. Hence, X is well-defined. **(iv)** Suppose *per absurdum* that X is ill-defined, entailing a directed path $Det_i \longrightarrow \cdots \longrightarrow Y$ in $\mathcal{G}_V$. Since $X \perp\!\!\!\perp Y|\mathbf{Z}$, some $Z_i \in \mathbf{Z} \subset \mathbf{O}$ is on that path. Then, Z_i is a child of an idiosyncratic shock in $\mathcal{G}^{IC}_{\{X,Z_i\}}$, contradicting (iv). Hence, X is well-defined. □

Proof of Proposition 4. Preamble: From the definition of ill-defined variable, and from $\mathbf{Det}_X \cap \mathbf{O} = \emptyset$, it follows that $\mathcal{G}_{\{X,Y\}}$ is $X \diamondsuit\!\!\longrightarrow Y$. **(i)** Under CFC and DD, the preamble implies $X \not\perp\!\!\!\perp Y$. **(ii)** By definition of IC representation, $\mathcal{G}^{IC}_{\{X,Y\}}$ contains at least one common shock to X, Y due to a latent determinant of X. **(iii)** If $\mathcal{G}_V$ contains a determinant of X not linked to Y by a directed path, then X is a child of an idiosyncratic shock (iii.a). If, on the contrary, all determinants of X are linked to Y by directed paths in $\mathcal{G}_V$, then X is not a child of an idiosyncratic shock. Additionally, given $X \perp\!\!\!\perp Y|\mathbf{Det}_X$, it follows that there is a set $\mathbf{S}$ of shocks on X's determinants, such that $X \perp\!\!\!\perp Y|\mathbf{S}$ (iii.b). □

References

Eberhardt, F. (2016). Green and grue causal variables. *Synthese 193*(4), 1029–46.

Hyvärinen, A., J. Karhunen, and E. Oja (2001). *Independent component analysis*. New York: John Wiley & Sons.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (Second ed.). Cambridge: Cambridge University Press.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of causal inference*. Cambridge, MA: MIT Press.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, prediction, and search* (Second ed.). Cambridge MA: MIT Press.

Spirtes, P. and R. Scheines (2004). Causal inference of ambiguous manipulations. *Philosophy of Science 71*(5), 833–45.

Woodward, J. (2016). The problem of variable choice. *Synthese 193*(4), 1047–72.

Paper to be presented at PSA2020: The 27th Biennial Meeting of the Philosophy of Science Association, November 19-22, 2020, Baltimore, MD.

Believe me, I can explain! Beware of inferences to the explanandum
(Draft as of March 5th, 2020)

David Colaço
Mississippi State University
Contact Email: dc2407@msstate.edu

Abstract: This paper addresses *inferences to the explanandum*: inferences from the premise that an explanandum is plausibly explained to the conclusion that this explanandum (or a claim or representation thereof) is adequate or true. These inferences consist in answering *why* or *how* something occurs to concluding *that* it occurs. Psychological research and scientific cases reveal that inferences to the explanandum are proposed by lay people and scientists, and some philosophical accounts permit these inferences. This is problematic, as pseudoscientific claims are often believed (or at least espoused) because they are plausibly explained. We should be skeptical of inferences to the explanandum.

1. Introduction

Philosophical accounts of explanation identify the targets to be explained (*explananda*) and the explanations for them (*explanantia*). There is also analysis of explanatory reasoning, which includes but is not limited to analyses of inferences to the best explanation. However, what is less discussed is the defensibility of an inference in the other direction: should we infer to an explanandum from plausibly explaining it?

While this inference is seldom discussed by philosophers, it is no less pressing. Consider that individuals' judgments of explananda are influenced by the explanations that are provided for them. A prima facie example of this is the "soy boy" effect. In this case, the explanandum is that soy consumption has "feminizing effects on men" (Messina 2010, 2095). Adherents claim that they believe in this effect in part because they can "explain" it: soy contains phytoestrogens, the ingestion of which causes changes in sex characteristics. This claim has the proper form for one account of explanation: it sketches a mechanistic relation between soy consumption and bodily changes. It also has some degree of empirical support: evidence corroborates that soy contains chemicals called 'phytoestrogens' and that hormonal estrogen can induce these changes in humans. However, there is no evidence that soy has this effect in humans (Messina 2010). This case of explaining to support an explanandum's belief-worthiness, whether involving a good faith inference or not, seems to have considerable rhetorical power. This raises a question. When provided a plausible explanation, should we infer that its explanandum is adequate or true? Should we (say) believe that the soy boy effect occurs based on its explanation?

This paper casts doubt on the idea that it is acceptable to infer that an explanandum is adequate or true from the provision of a plausible explanation for it – what I call an *inference to the explanandum*. The provision of an explanation serves as such a bad reason to infer that its explanandum is true or adequate that, when explaining stands as the sole basis for inferences regarding its explanandum, one should infer that this explanandum is explainable but should not infer that is true or adequate. My doubt is not based on deficiencies of so-called "plausible explanations"; I remain skeptical even when explanations match the form of mechanistic models (Craver 2007) and have empirical support.

In Section 2, I introduce inferences to the explanandum and define 'plausible explanation.' In Section 3, I discuss different conclusions that may be inferred about an explanandum from explaining it. In Section 4, I argue that we should be skeptical of inferences to the explanandum. This is because (1) explanantia do not provide evidential support for their explananda, (2) the evidence for explanantia need not transfer to explananda, and (3) the idea that "bona fide" explaining entails that explananda are adequate or true does not warrant inferring to the explanandum. Inferences that conclude with the explainability of explananda are, by contrast, acceptable. I conclude by discussing how inferences to the explanandum illustrate a drawback of lay, scientific, and philosophical predilections for explanation: unsupported claims or outright pseudoscience can garner legitimacy from epistemically suspect inferences and subsequently can be exploited for unscrupulous aims.

2. Inference to the Explanandum

An inference to the explanandum consists in inferring from the premise that a given explanandum is plausibly explained to the conclusion that this explanandum (or a claim or representation thereof) is adequate or true. This amounts to concluding that an explanandum is belief worthy (or a cognate epistemic attitude) from explaining it.[1] A plausible explanation consists in a claim or representation that has the form for a philosophical account of scientific explanation as well as some degree of empirical support for its content. For example, if I infer from the premise that I can plausibly explain *why* the soy boy effect occurs to the conclusion *that* it occurs, I infer to the explanandum. This inference juxtaposes an inference to the best explanation, or an inference "from the premise that a given hypothesis would provide a 'better' explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true" (Harman 1965, 89).

Let me detail my characterization. First, inferences to the explanandum are inferences about the adequacy or truth of the explanandum: they are not inferences concluding something else about an explanandum, such as it being predictive or explainable. Second, inferences to the explanandum are inferences about specific explananda: if the explanandum

[1] I include "a claim or representation" to reflect that ontic accounts of explanation take explananda to be in the world and therefore not candidates for being true or adequate (Halina 2017). I include "adequate" to accommodate accounts of explanation that do not construe explananda as candidates for being true.

is the soy boy effect, for example, it is an inference about this effect as it is characterized by its adherents. It is not a broader, unspecific inference about there being *some* explanandum that is explained. Third, unlike inferences to the best explanation, inferences to the explanandum are not about picking the best explanandum from a set of explananda. Rather, it is an inference about the particular explanandum and no other.

Motivation for this paper comes from research on *explanation effects*. This effect is exemplified by Ross and colleagues, who tested whether or not "the process of explaining an event increases its subjective likelihood for the perceiver" in subjects who were told that the events were fictious (1977, 818), from which they conclude "that providing an explanation for an event substantially increases the subjective likelihood of the occurrence of the event" (1977, 825-826). Explanation effects are discussed by Lombrozo, who notes that "psychological findings suggest that the mere existence of an explanation can influence the probability assigned to an explanandum," and "explaining a hypothetical outcome… increases the subjective probability of that outcome" (2011, 545). Some psychologists present concerns about "this reliance on explanatory considerations" in reasoning (Lombrozo 2011, 546). For instance, Kuhn notes that "people… depend on explanations that allow their claims to 'make sense'," but she emphasizes that explanations "lead to overconfidence, they inhibit examination of alternatives, and, most seriously, they may be false" (2001, 1).

What do inferences to the explanandum have to do with science? First, some accounts of explanatory reasoning permit these inferences. For example, Thagard's explanatory coherence account, according to which "we should accept propositions that cohere with our

other beliefs" (1989, 436), indicates that "a hypothesis coheres with what it explains," and "we should accept or reject propositions based on their overall coherence with one another" (2006, 142). Explanatory coherence accounts "for a wide range of explanatory inferences," including, it would appear, inferences to the explanandum (Thagard 1989, 435). Inferences to the explanandum are also indirectly supported by Hempel's conception of "explanatory relevance": explanatory information "affords good grounds for believing that the phenomenon to be explained did, or does, indeed occur" (1966, 48). While Hempel accounts for explanation rather than reasoning, explanatory relevance supports the idea that, in general, an explanation's quality should be measured in terms of the support it lends to believing in its explanandum. This idea provides some justification for inferring to the explanandum.

Second, scientists' judgments of explananda are influenced by explaining them. Scientists, on occasion, take mechanisms to "add weight" to what these mechanisms explain: "an analogy would be that we are more certain that we actually went to the moon if we understand the small scale step-by-step mechanisms that explain how we got there" (Patihis 2018, 375). Psychologists provide reason to worry that inferences like this are not uncommon. When studying explanatory reasoning in science students, Masnick and Zimmerman claim that because "individuals are more likely to believe an empirical finding if there is a theory or explanation for that finding, … it is unsurprising that the presence of explanatory information would increase perceptions of how important or interesting a topic is" (Masnick and Zimmerman 2009, 35). This tendency to infer to the explanandum seems

greatest with mechanistic explanations (Craver 2007), which is why I focus specifically on mechanistic explanation.[2]

3. Inferences about Explananda

Given the wealth of evidence that suggests that inferences to the explanandum occur amongst lay people, I focus on explanatory inferences amongst scientists. While these occur, not all inferences regarding explananda fit my characterization of an inference to the explanandum. Therefore, it is prudent to disentangle these inferences. I discuss two cases. Each case involves the provision of a plausible explanation. The first case is the plausible explanation of a controversial explanandum. The second case is the plausible explanation of a putatively unexplainable target.

3.1. Plausibly Explaining a Controversial Explanandum

What happens when researchers infer that a controversial target phenomenon occurs from its plausible mechanistic explanation? Memory transfer is one of the most notorious cases of a controversial phenomenon to have been alleged to occur in the history of science. This phenomenon was characterized as the transfer of memories from one organism to another via the transfer of tissue. A proponent of memory transfer, Ungar, defended that this

[2] Equivalent concerns can be devised for other accounts of scientific explanation, but this is beyond the scope of this paper.

alleged phenomenon occurs because he could provide a plausible mechanistic explanation for memory transfer's occurrence (Colaço 2018, 37).

Ungar supported that he transferred memories via chemical injection by claiming that he had isolated its chemical substrate, which he called "scotophobin" (1974, 599). With this substrate isolated, Ungar schematized it as a component of a mechanism for memory transfer. He claims that it is "widely held, in spite of the inadequacy and controversial nature of the evidence, that some sort of molecular coding would be the most likely explanation of learning," and he claims that "built-in pathways… can be founded, and a fully developed molecular coding system which maintains the synaptic connections between the neurons of each functionally related pathway" can be schematized (Ungar 1968, 222). Ungar claims that "this peptide, called scotophobin, was synthesized and distributed to a number of laboratories, which *confirmed* its dark-avoidance inducing effect," which suggests that schematizing this mechanism and providing evidence for it is reason to believe in memory transfer (1974, 599, my emphasis). This is a plea to infer to the explanandum from this mechanistic model.

Ungar's contemporaries were skeptical about memory transfer. One skeptic, Stewart, claims that Ungar's "conclusions are more likely false than true," though he notes that the "synthesis of the pentadecapeptide [scotophobin] is essentially sound" (Stewart 1972, 209). Like other skeptics, Stewart appears to have accepted Ungar's mechanistic model as plausible insofar as he accepted that components of the model were empirically supported. Nevertheless, Stewart and others claimed that Ungar provided insufficient evidence to

believe in the explanandum. It would appear that the "inadequacy and controversial nature of the evidence" was sufficient to keep skeptics skeptical of the alleged explanandum. The skeptics won out: the memory transfer project collapsed, despite Ungar plausibly explaining this explanandum (Colaço 2018, 37).

3.2. Plausibly Explaining an "Unexplainable" Target

What about cases in which an explanation is provided for a putatively unexplainable target? One alleged inference of this character is in the case of continental drift. Historians have argued that continental drift, or the movement of continents over time, was rejected by geologists in part because it was not explainable. It was not until the provision of an "adequate causal mechanism" in modern plate tectonics, this argument continues, that continental drift was accepted (Oreskes 1988, 312). Laudan claims that "the problem with drift was not that there was no known mechanism or cause, but that any conceivable mechanism would conflict with physical theory" (1978, 230).

Oreskes challenges these historical claims. She claims that "a theory of drift did not fail for lack of a mechanism," highlighting that researchers had provided explanations that were largely rejected (Oreskes 1988, 331). Oreskes argues that "the most likely cause of the rejection of continental drift was the evidence put forward to support it" (1988, 332), though some geologists rejected it due to the "lack of an adequate driving force for drift" (1988, 334). This suggests that better evidence for continental drift was desired, though there were concerns about its explanation as well. That being said, Oreskes argues that the acceptance of

drift came from "not the elucidation of… the mechanism by which they occur, but by the availability of a new kind of evidence" (1988, 346).

On either Laudan or Oreskes' construal of the case, plausibly explaining continental drift resulted in researchers investigating it. However, three features differ the continental drift case from the scotophobin case. First, many researchers argued that continental drift was unexplainable, while skeptics accepted that memory transfer was explainable. Second, no researcher suggests that explaining drift "confirms" its occurrence, as Ungar argued. If something was inferred from the mechanistic models of plate tectonics, it was not the adequacy or truth of the explanandum. Third, even if an inference about the explanandum occurred in the drift case, this inference was wrapped into debates about what counts as evidence for the explanandum. Neither historical construal suggests that continental drift was believed solely based on its plausible explanation.

4. Whither Inference to the Explanandum?

The cases in Section 3 show that there are at least two distinct inferences one might make about an explanandum from its plausible explanation. The scotophobin case matches what I characterize as an inference to the explanandum. By contrast, the continental drift case has two differences: it does not involve inferring that the explanandum is adequate or true, and it involves inferences about targets that were previously considered to be unexplainable. It is these differences that make the latter sort of inference acceptable, while the former sort – inferences to the explanandum – are epistemically suspect.

Inferences to the explanandum, meaning the inference from a plausible explanation to the truth or adequacy of its explanandum, are inferences about which we should be skeptical. This is because these inferences, despite their apparently compelling character, do not empirically support the explanandum's truth or adequacy. For one, explanantia are not evidence for explananda. Following on Kuhn's insight (2001), explaining alone provides no evidence for the explanandum, even if this "explaining" matches the form of an account of explanation and coheres with the characterized explanandum as specified on the explanatory coherence account. The study from Ross and colleagues illustrates this limitation of explaining: one can plausibly explain explananda that are known to be fictitious without making these explananda any less fictitious. Thus, if one argues that an explanation warrants believing in its explanandum, it will not be through a plausible explanation serving as evidence for this explanandum.[3]

Perhaps the empirical support for plausible explanations transfers to the explanandum. After all, part of what makes explanations plausible is that they have some degree of empirical support. This support, one might argue, is also support for the explanandum. The explanatory coherence account corroborates this idea. If the explanans

[3] Even the explanatory coherence account indicates that evidence is stronger than explanation: "a proposition describing the results of observation has a degree of acceptability on its own," as "it can stand on its own more successfully than can a hypothesis whose sole justification is what it explains" (Thagard 1989, 437-438).

and explanandum cohere, and the explanans and its evidence cohere, then the explanandum and this evidence cohere. Therefore, on this account, evidence that coheres with an explanation also supports its coherent explanandum.

However, evidence for a plausible explanation need not be transitive. Even when components of an explanation are empirically supported, this evidence may be neutral to the truth or adequacy of its explanandum. The evidence that supports the identification of scotophobin does not provide a test of the occurrence of the explanandum in this case: the peptide may underwrite a distinct explanandum phenomenon. The "soy boy" case also illustrates this lack of transitivity: neither the fact that phytoestrogen is in soy nor the fact that hormonal estrogen has this effect is evidence that phytoestrogens function like human hormonal estrogen.

What about cases where the evidence does transfer from explanans to explanandum? This no more supports the acceptability of inferences to the explanandum than when evidence does not transfer. If the evidence for an explanans is transitive, then this evidence confirms the explanandum. The explanation merely serves as a means to connect the explanandum with this evidence. The idea of transitive evidence may go some way in explaining the putative successes of reasoning strategies like the explanatory coherence model: so long as there is evidence that supports the explanandum, this evidence, the explanandum, and its explanans cohere, and the explanandum is supported as a result. However, this is not an inference to the explanandum.

Perhaps the acceptability of inferences to the explanandum is independent of evidential considerations. One might have the intuition that something does not count as "bona fide" explaining if its explanandum is not true or adequate. This intuition supports the idea that the scotophobin case involves unacceptable inferences because of deficiencies of the so-called "explanation" rather than issues with inferences to the explanandum in general. This intuition hints at two ideas. First, plausible explanations fail to count as bona fide explaining, so my examples have no relevance to adjudicating the acceptability of inferences to the explanandum, and it is a mistake to refer to them in terms of 'explanation,' 'explanans,' or 'explanandum.' Second, because the adequacy or truth of the explanandum is sin qua non for bona fide explaining, inferences to the explanandum that involve bona fide explanations are acceptable because of this relation. Thus, the reader may be sympathetic to the idea that, regardless of whether or not they are called 'explanations,' one cannot explain an explanandum that is false or inadequate. This intuition may lead the reader to be doubtful of the explanatory merit of what I call 'plausible explanations.'

While this intuition may be compelling, we should dismiss it. For one, endorsing this intuition comes at the cost of descriptive adequacy. The majority of explanations in science likely are not "bona fide" in the relevant sense. Even if bona fide explaining entails that the explananda are true or adequate, this does not entail that these inferences are acceptable in real explanatory practices: even if we assume that only true or adequate explananda are genuinely explained, many times in practice, these explananda turn out to be false or inadequate. Explanatory claims made by scientists also should be taken seriously because,

whether one wants to count them as deficient or not, the explanations put forward in the scotophobin and soy boy cases match the form of the mechanistic account and have empirical support. This suggests that this intuition is at odds not only with how explanatory claims are employed in science but also with philosophical accounts of scientific explanation.

However, there is a deeper issue with this intuition. Even if we accept that, in principle, bona fide explanations are explanations of true or adequate explananda, this alone does not warrant an inference to the explanandum. This is because it leaves open the question of how we come to know that an explanation is bona fide, and we thus are permitted to infer from it to the truth or adequacy of the explanandum. This epistemic issue speaks to why Craver suggests that characterizing "the [explanandum] phenomenon correctly and completely is a crucial step" in developing explanatory models (Craver 2007, 128). Mechanists like Craver take correct models of explanantia to depend on correct characterizations of explananda, and not the other way around. This emphasis on settling the explanandum before moving on to the explanans is typical in philosophical analysis of explanation: "the event or phenomenon in question is usually accepted as a matter of fact," as "in an explanation the purpose of the explanans is to shed light on, or make sense of, the explanandum event – not to prove that it occurred" (Hurley 2014, 21). This casts doubt on the idea that one could identify bona fide explaining, let alone infer the adequacy of its explanandum from it, without first correctly characterizing the explanandum. If a correct characterization is required, then an inference to the explanandum is, at best, redundant.

Inferences to the explanandum are not justifiable on evidential grounds, and they are not justifiable on grounds of the relations between explanantia and explananda. Is this enough to rule them out as unacceptable? Perhaps these inferences are acceptable for a reason that I have failed to identify, but these deficiencies support my general skepticism about inferences to the explanandum. My conclusion is akin to the skepticism many philosophers have about inferences to the best explanation. Those concerned with inferences to the best explanation claim that there is an "expectation that one should establish the reality of one's posits on non-explanatory grounds" when determining the belief-worthiness of these posits (Novick and Scholl 2020, 7). If one cannot establish these posits for reasons aside from explanatory power, one should be skeptical that they are belief worthy. This parallels my skepticism of inferences to the explanandum: without establishing one's explanandum independently of explaining it, one should be skeptical of it.

If the sort of inferences exemplified by the scotophobin case are ones about which we should be skeptical, what does this mean for the sort of inferences exemplified by the continental drift case? The answer is simple: these inferences are acceptable because one infers that the explanandum is *explainable* based on its plausible explanation. This is a straightforward inference from the premise that a target is plausibly explained to the conclusion that it can be explained. And, if one adopted an epistemic stance towards a target based on its perceived unexplainability, then one should change one's epistemic stance towards that target once one infers that it is explainable.

Inferring that a target is explainable and inferring that it is true or adequate are not equivalent. As we saw in the scotophobin case, skeptics of memory transfer did not deny that the alleged explanandum was explainable. Nonetheless, they were skeptical of memory transfer, and they ultimately rejected it in light of deficient evidential support. In the continental drift case, part of the reason some geologists rejected drift prior the mid-20th century, Oreskes argues, was tied to them thinking that there was no possible explanation for it, in addition to their assessment of the quality of evidence put forward to support it. Thus, in this case, the explanandum was initially rejected (at least in part) because researchers at the time were skeptical about its explainability.

If they are not inferred to be true or adequate, then how should we conceive of these targets that have been inferred to be explainable? If a claim about a target is not rejected, but researchers are not yet in a position to determine its truth or adequacy, then it is at least pursuit worthy in the sense that it is worth investigating "to the extent that it can be shown to have a promising potential for contributing… [to] scientific knowledge" (Šešelja and Straßer 2014, 3115). Thus, researchers can investigate this explanandum with the aim of producing evidence for or against its truth or adequacy. They can, for example, generate predictions about this explanandum. This is not an inference to the explanandum. Rather, it supports the idea that evidence is needed to assess the truth or adequacy of an explanandum.

While inferences about explainability may help orient us towards new investigations, there are examples of active targets of scientific investigation that have yet to be plausibly explained. For example, there is the placebo effect, which is accepted despite it lacking an

explanation (Price et al. 2008). This idea is not foreign to geology either. Oreskes notes that "many empirical scientific phenomena have been accepted before their causes were known," highlighting that unexplained targets are not immediately rejected for being unexplained (1988, 324). The idea that one need not prove targets are explainable prior to investigating them should not be surprising, given the order in which things occur on the mechanistic account: it is critical to correctly characterize explananda phenomena in order to correctly model their mechanistic explanations. Of course, cases like the placebo effect lack an explanation at this time, but they are not considered to be unexplainable. I take no stand on the conditions under which something should be judged to be unexplainable. What matters is that, often in science, targets that were considered to be unexplainable are explained, and inferring from the premise that a target is plausibly explained to the conclusion that it is explainable is an acceptable inference. This is not an inference to the explanandum.

5. Conclusion

It is important for philosophers to acknowledge that explanations are compelling *to a fault*, and our predilection for explanation is not an unequivocally good thing. I have provided reason to be skeptical of the idea that the explanations that we find in science can serve as the basis for inferring that their explananda are adequate or true. My conclusions cast accounts of explanatory reasoning like the explanatory coherence account in question, given that it supports inferences to the explanandum. Further, I have shown that inferences to the explanandum ought to be distinguished from inferences about the explainability of

explananda, the latter of which are acceptable. Overall, I have shown that it is unwise to employ explanations for epistemic tasks for which they are ill suited.

How do inferences to the explanandum fit into the philosophical discussion of explanation? The overarching reason to address these inferences is not that they are accepted by philosophers, even if this is the case. Rather, the reason is that they are proposed in both scientific and lay reasoning. This fact highlights a serious concern about the rhetorical strength of explanation claims in scientific as well as lay reasoning about science or pseudoscience, even when the inferences made from these claims are epistemically suspect and possibly put forward in bad faith. Whether knowingly or not, individuals can exploit our explanatory predilections to legitimize pseudoscience and achieve unscrupulous aims supported by the espousal of this pseudoscience, as appears to be the case with the alt-right espousal of "soy boys." We must ask if the philosophy of science inadvertently contributes to this situation by focusing on explanation while eliding discussion of its limited epistemic implications. For this reason, and despite the rhetorical strength explaining has in science and everyday life, we should be skeptical about changing our epistemic stance towards what is being explained when an explanation is provided for it. Further, we should be suspicious about the provision of plausible explanations as the basis for inferences about the belief-worthiness of controversial research targets in scientific and lay discourse, particularly in cases where evidence for these targets is deficient.

References

Colaço, David. 2018. "Rip it up and start again: The rejection of a characterization of a phenomenon." *Studies in History and Philosophy of Science Part A*, *72*, 32-40.

Craver, Carl. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Halina, Marta. 2017. "Mechanistic explanation and its limits." In *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 213-224). Routledge.

Harman, Gilbert. 1965. "The inference to the best explanation." *The philosophical review*, *74*(1), 88-95.

Hempel, Carl. 1966. *Philosophy of Natural Science*. Prentice Hall: New Jersey.

Hurley, Patrick. 2014. *A concise introduction to logic*. Nelson Education.

Kuhn, Deanna. 2001. "How do people know?" *Psychological science*, *12*(1), 1-8.

Laudan, Rachel. 1978. "The recent revolution in geology and Kuhn's theory of scientific change." In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1978, No. 2, pp. 227-239). Philosophy of Science Association.

Lombrozo, Tania. 2011. "The instrumental value of explanations." *Philosophy Compass*, *6*(8), 539-551.

Masnick, Amy, and Corinne Zimmerman. 2009. "Evaluating scientific research in the context of prior belief: Hindsight bias or confirmation bias?" *Journal of Psychology of Science and Technology*, *2*(1), 29-36.

Messina, Mark. 2010. "Soybean isoflavone exposure does not have feminizing effects on men: a critical examination of the clinical evidence." *Fertility and sterility*, *93*(7), 2095-2104.

Novick, Aaron, and Raphael Scholl. 2020. "Presume it not: True causes in the search for the basis of heredity." *British Journal for the Philosophy of Science 71*(1), 59-86.

Oreskes, Naomi. 1988. "The rejection of continental drift." *Historical Studies in the Physical and Biological Sciences*, *18*(2), 311-348.

Patihis, Lawrence. 2018. “The Historical Significance of the Discovery of Long-Term Potentiation: An Overview and Evaluation for Nonexperts.” *American Journal of Psychology*, *131*(3), 369-380.

Price, Donald, Damian Finniss, and Fabrizio Benedetti 2008. “A comprehensive review of the placebo effect: recent advances and current thought.” *Annual Review of Psychology*, 59, 565-590.

Ross, Lee, Mark Lepper, Fritz Strack, and Julia Steinmetz. 1977. “Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood.” *Journal of Personality and Social Psychology*, *35*(11), 817.

Šešelja, Dunja, and Christian Straßer. 2014. “Epistemic justification in the context of pursuit: A coherentist approach.” *Synthese*, *191*(13), 3111-3141.

Stewart, Walter. 1972. “Comments on the chemistry of scotophobin.” *Nature*, *238*(5361), 202.

Thagard, Paul. 1989. “Explanatory coherence.” *Behavioral and brain sciences*, *12*(3), 435-467.

Thagard, Paul. 2006. “Evaluating explanations in law, science, and everyday life.” *Current Directions in Psychological Science*, *15*(3), 141-145.

Ungar, Georges. 1968. “Molecular mechanisms in learning.” *Perspectives in biology and medicine*, *11*(2), 217-232.

Ungar, Georges. 1975. “Molecular coding of memory.” In *Minireviews of the Neurosciences from Life Sciences* (pp. 459-468). Pergamon.

This is a very early draft accepted for presentation at PSA 2020. Please don't cite; criticisms, questions, or requests for a revised draft would be most welcome (devin.curry@mail.wvu.edu).

g as bridge model

Devin Sanchez Curry

Abstract: g*—a statistical factor capturing strong intercorrelations between individuals' scores on different IQ tests—is of theoretical interest despite being a low-fidelity model of both folk psychological intelligence and its cognitive/neural underpinnings.* g *idealizes away from those aspects of cognitive/neural mechanisms that are not explanatory of the relevant variety of folk psychological intelligence, and idealizes away from those aspects of folk psychological intelligence that are not generated by the relevant cognitive/neural substrate. In this manner,* g *constitutes a high-fidelity bridge model of the relationship between its two targets, and thereby helps demystify the relationship between folk and scientific psychology.*

1. Introduction

Psychometric *g* is a statistical factor that captures the remarkably strong positive intercorrelations between all of any given individual's scores on different IQ tests and subtests. There are many varieties of IQ subtest, probing verbal ability, analogical reasoning, mathematical ability, pattern-matching ability, and so on. The first great finding of the IQ-testing tradition is that subjects who do better than most people on any given one of these subtests are also likely to do better than most people on any of the others (Mackintosh 2011). *g* is thus commonly considered a statistical distillation of what all IQ subtests measure in common. The second great finding of the IQ-testing tradition is that *g* is predictively fecund—among psychological constructs, only conscientiousness competes with *g* as a predictor of educational attainment, job complexity, socioeconomic status, and other prominent measures of success in life (Gottfredson 1997). Nevertheless, experts are divided about its theoretical interest.

Some skeptics deny that *g* measures anything more theoretically interesting than the ability to do well on IQ tests, but most intelligence researchers assume that *g* is a very good model (if not a direct measure) of something of theoretical interest. (Researchers variously refer to the phenomenon modelled by *g* as 'general intelligence', 'the positive manifold', or just 'the *g*-factor'.) Non-skeptics tend to emphasize one or the other of two target systems purportedly modeled by *g*. According to some intelligence researchers, *g* is a model of folk psychological intelligence—the personal-level capacity that ordinary folks are talking about when they call somebody smart. According to others, *g* is a model of the cognitive or neural substrates of that capacity.

I'll argue that *g* is of theoretical interest despite being a low-fidelity model of each of these targets. I'll begin by assuming that *g* isn't a very good measure of folk psychological intelligence. I'll then argue that it isn't a very good measure of what's going on in the brains or cognitive systems of (un)intelligent people, either. I'll go on to argue that *g* is nevertheless explanatorily important insofar as it idealizes away from those aspects of the relevant neural/cognitive substrates that aren't explanatory of the relevant variety folk psychological intelligence, and idealizes away from those aspects of the relevant variety of folk psychological intelligence that aren't generated by the relevant neural/cognitive substrates. In that manner, *g* constitutes a high-fidelity 'bridge model' of the relationship between its two distinct targets, and thereby helps demystify the relationship between folk psychology and scientific psychology.

2. *g* isn't a very good measure of folk psychological intelligence

Elsewhere (Curry forthcoming), I have argued for an interpretivist account of folk psychological intelligence inspired by Ryle's (1945) analysis of intelligence-talk and Dennett's (1991) notion of real patterns detected from the intentional stance. On my account, to be intelligent (in the sense invoked in folk psychological practices) is to be comparatively good at solving intellectual problems that an interpreter deems worth solving. In short: you're intelligent if you behave (in ways that folks deem smart) more successfully than other people, and you're unintelligent if you behave (in ways that folk deem smart) less successfully than other people. Since the extant empirical evidence indicates that different lay interpreters, both between and within cultures, deem different intellectual problems worth solving (and, indeed,

deem different problems to count as *intellectual* problems), it follows from my definition that what it is to be intelligent varies alongside the lay interpreters in question.

As Sternberg and Grigorenko (2004) and their collaborators in cross-cultural psychology have extensively documented, *g* tracks some—but not all—of the varieties of intelligence that have emerged in relation to folk psychological practices around the globe. In particular, *g* is plausibly a decent model of a variety of intelligence that became extremely salient in the folk psychological discourses of some WEIRD—Western, Educated, Industrialized, Rich, Democratic (Henrich et al. 2010)—contexts in the 20th century, but which is much less salient in other cultural contexts. However, skeptical philosophers and psychologists have provided serious reasons to doubt that *g* is a *very* good measure even of the varieties of folk psychological intelligence that have emerged, alongside IQ testing itself, within WEIRD contexts (Block & Dworkin 1974). So I'll henceforth assume that *g* isn't a very good measure of what folks are talking about when they talk about intelligence in everyday life: it doesn't straightforwardly measure intelligence as conceptualized in WEIRD, IQ-test-influenced settings, and it flat-out fails to measure intelligence as conceptualized in many other settings. Nevertheless, my account of folk psychological intelligence leaves open the possibility that *g* is a great measure of the neural or cognitive underpinnings of what folks are talking about when they talk about intelligence.

3. *g* isn't a very good measure of cognitive or neural functioning

Several prominent psychologists and cognitive neuroscientists are increasingly optimistic about unearthing a particular neural or cognitive mechanism (or set of mechanisms)

that is fully responsible for the comparatively superior (or inferior) capacity measured by *g*, and thereby discovering intelligence squarely in the brain or cognitive system. I think their optimism about reduction is misplaced. To substantiate my pessimism, let's go through a few prominent recent attempts to reduce intelligence to its neural or cognitive substrates.

3.1. Neural correlates

Jensen (2006: *ix*), in a refinement of Spearman's original speculation that *g* measured a kind of "mental energy", influentially interpreted *g* as an indirect "measurement of cognitive speed" which could be more directly measured via reaction time paradigms which correlate strongly with *g*. Because of this correlation, Jensen was convinced that "intelligence is the periodicity of neural oscillation in the action potentials of the brain and central nervous system" (2011: 173). In other words, intelligence is nothing more and nothing less than the frequency of brainwaves, and IQ testing provides a good (if indirect) measure of this physical feature of the brain. Jensen's simple reductionist theory of intelligence hasn't held up in the light of PET and fMRI research in cognitive neuroscience. For one thing, cognitive neuroscientists have demonstrated that a higher frequency of brainwaves isn't actually straightforwardly correlated with greater neural processing power; nor is any other particular pattern in the frequency of brainwaves (Haier 2017). It turns out that, despite Jensen's best efforts, Spearman's notion of mental energy has no neural referent. Nevertheless, more empirically adequate neurological theories of intelligence have risen in Jensen's theory's stead.

The best developed among them—Jung and Haier's Parieto-Frontal Integration Theory—goes a long way towards identifying the neural correlates of the cognitive processes

recruited when people take IQ tests. There is surely something to Jung and Haier's suggestion that the efficient integrated operation of a parieto-frontal sense-remember-judge-act network underlies the variety of intelligence purportedly measured by *g*. (It plausibly partially underlies many other varieties of folk psychological intelligence as well.) But Jung and Haier have no proposal as to the cause of this efficiency, which could theoretically stem from a wide variety of sources, only some of which could be plausibly construed as the incarnation of intelligence in the brain. (More on alternative sources of efficiency anon.) Indeed, in responding to critics, Jung and Haier back off of the claim to have provided a reductionist theory of the positive manifold modelled by *g*, and instead insist only that "in our view, it is still too early to rule out a neural basis for a general factor of intelligence independent of a neural basis for specific cognitive abilities" (2007: 176). In other words, Jung and Haier insist that it is possible that the parieto-frontal efficiency which underlies successful IQ test-taking is generated by intelligence qua mechanism in the brain. They claim to have located that mechanism in a reasonably delimited parieto-frontal network. But, in the end, they make no claim to have identified the mechanism itself.

Localization isn't nearly enough to ground reduction. If researchers hope to reduce intelligence to a neural—or, failing that, cognitive—state or process, then they'll have to identify a candidate mechanism that produces that state or carries out that process. To be fair, some researchers have done just that. The most plausible candidate mechanism currently on offer is working memory capacity.

3.2. Working memory

Cognitive scientists use the term 'working memory' to refer to "a domain-general resource that enables representations to be actively sustained, rehearsed, and manipulated for purposes of reasoning and problem solving" (Carruthers 2014: 12). When you rehearse a phone number in your head while looking for a piece of paper to scribble it down on, you're using your working memory. Working memory *capacity* is a common measure both of how much information can be maintained in working memory and of how well that information can be processed. Research on working memory capacity has increasingly shown that it is a critical component in much—perhaps even most—complex cognition. As such, researchers have become increasingly interested in the hypothesis that intelligence can be explained largely in terms of—perhaps even be reduced to—working memory capacity.

This hypothesis makes some intuitive sense: solving puzzles almost always involves actively sustaining and manipulating information. And, at first glance, the evidence in favor of reducing intelligence to working memory capacity is impressive. When you give somebody both an IQ test and a test of working memory capacity, the two resulting scores correlate positively. In particular, working memory capacity and 'fluid *g*'–the factor capturing how well people do on IQ tests that are designed to focus on pure reasoning abilities, as opposed to reasoning that makes use of what the reasoner knows—tend to have a correlation somewhere between .6 and .8 (Carruthers 2014); that is a very strong correlation (indeed, that range is only slightly lower than the range of correlations that made *g* such an important finding in the first place). Moreover, much of the parieto-frontal network that Jung and Haier identify as the neural correlate of *g* has also been shown to be active in working memory (Deary et al. 2010).

Finally, there is some evidence that increases in working memory capacity yield increases in fluid *g* (Jaušovec & Jaušovec 2012).

On the other hand, there is also evidence that cuts against reduction. Working memory capacity, while quite domain-general, is nevertheless more domain-specific than fluid *g*: it correlates more with tests of verbal ability than with tests of spatial ability, for instance. And working memory's contribution to performance on tests of fluid *g* seems to be independent of the respective contributions of associative learning and information processing speed (Mackintosh 2011: 154–155). So there is good reason to doubt that working memory is the sole cognitive underpinning of fluid *g*. Moreover, there is some good reason to doubt that working memory is a cognitive underpinning of intelligence at all: some of the researchers responsible for discovering the correlations between fluid *g* and working memory capacity have argued that the two are explanatorily distinct phenomena that are nevertheless strongly correlated because they share a common underpinning (Shipstead & Engle 2018). But for my purposes we can set these complex questions about the weight and interpretation of the extant evidence aside. My argument against the reduction of intelligence to working memory capacity is at once more abstract and more straightforward: my argument rests on the premise that reduction would add nothing to—and indeed subtract something from—our understanding. In particular, reduction would hinder our understanding of intelligence while adding nothing to our understanding of how cognitive systems work.

With regard to the latter: working memory capacity is already a reasonably well-defined construct that measures the operations of a central and reasonably well-delimited (albeit complex and distributed) cognitive subsystem, and thereby plays a clear explanatory role in

cognitive science. Stipulating that this construct is a measure of intelligence—without making any concrete suggestions for how that stipulation should change our understanding of working memory or the functioning of cognitive systems more generally—does nothing to enhance its explanatory power. Thus, reduction is justified in this case only if it sheds light on the phenomenon being reduced.

But reduction to working memory capacity can only obfuscate intelligence. Even granting that IQ tests measure intelligence well, any attempted reduction of intelligence to working memory capacity will hinder our understanding of intelligence in at least two respects.

First, working memory capacity is super highly correlated, not with *g*, but only with one of its component factors, fluid *g*, which is derived from minority subset of IQ tests. Most IQ tests also measure other component factors, including most prominently 'crystallized *g*': the factor capturing how well people do on IQ tests that are designed to focus on reasoning that makes use of what the reasoner knows. The calling card of plain old undifferentiated *g* is that there are strong intercorrelations between how well people do on *all* IQ tests—including relatively pure tests of fluid *g*, relatively pure tests of crystallized *g*, and a wide range of hybrids. By my lights, the heterogeneous nature of the positive manifold should be telling when it comes to constructing a theory of intelligence: the fact that both fluid *g* and crystallized *g* are statistical components of undifferentiated *g* intriguingly mirrors the fact that folk psychological conceptions of intelligence across cultures tend to invoke both fluid reasoning and the use of crystalized knowledge (Sternberg & Grigorenko 2004). Meanwhile, the correlation of crystallized intelligence and working memory capacity, like the correlation of undifferentiated *g* and working memory capacity, is somewhere between .3 and .6 (Mackintosh 2011)—the two are

clearly importantly related, but it is equally clear that a direct reduction of one to the other won't be in the offing.

Of course, it is possible that fluid *g* captures the essence of *g* (and, by extension, of folk psychological intelligence), and that crystallized *g* is more noise than signal. Indeed, the IQ tests with the highest *g*-loadings—that is, that correlate most strongly with *g* itself—tend to be tests of fluid intelligence (like Raven's Progressive Matrices). But there are problems even with reducing fluid *g* alone to working memory capacity. As Block and Dworkin (1974) have argued, there is a strong case to be made that fluid *g* measures personality, motivation, and temperament to a large degree—for example, it seems to measure ambition, patience, and test-wiseness as well as pure reasoning capacity—and these characteristics aren't plausibly reduced to working memory capacity. Indeed, my account of folk psychological intelligence suggests that these character traits measured by fluid *g* are rightly taken to be part and parcel of intelligence: intelligence is the capacity to solve intellectual problems comparatively well, and solving problems better than one's peers takes grit as well as wits (Dweck & Bempechat 1983).

Nevertheless, I recognize that there remains a reasonable case to be made that, by shedding inessential character traits, working memory capacity distills the essence of fluid *g*, which itself, by shedding crystallized knowledge, distills the essence of undifferentiated *g*. But even if this is the case, a second pitfall awaits the attempt to reduce fluid *g* to working memory capacity (and indeed any attempted reduction of a psychometric kind to the workings of a cognitive mechanism).

Even if working memory capacity is *the* essential cognitive underpinning of intelligence, *g* isn't a very good model thereof. That's because the *g*-factor is, by its very nature, *comparative*—

it is an inter- (rather than intra-) individual construct that measures how somebody does on IQ tests relative to other people in their age-cohort. It doesn't measure how smart somebody is on a ratio scale; it measures only how much better or worse they perform than the average IQ-test-taker. *g* thus can't directly measure an intrinsic characteristic of any individual's mind, whereas we already have reliable ways of measuring working memory within a single individual on a ratio scale. (To my mind, this is a salutary fact about *g*, since on my definition folk psychological intelligence is also constitutively comparative.) As Borsboom and colleagues (2009) have pointed out, absent a theory of how to bridge differential and cognitive psychology, "intelligence dimensions like the g-factor can't be understood on the basis of between-subject data as denoting mental ability qua within-subject attribute." Fluid *g* couldn't be comprehensibly reduced to working memory capacity absent a grand unifying theory of how constitutively comparative capacities relate to intrinsic cognitive mechanisms.

In contrast, it bears repeating that cognitive psychologists already have a decent theoretical understanding of the mechanics of working memory capacity in its own right, not to mention reliable instruments that measure it on a ratio scale. And theorists can give working memory capacity due emphasis as a cognitive underpinning of intelligence without making an attempt at reduction. If my argument holds water, then, in attempting reduction, nothing new is learned, some of the plausibly explanatorily salient dimensions—crystallized intelligence and, arguably, other characteristics—of both folk psychological intelligence and *g* are erased, and an important distinction—between the intrapersonality of the cognitive mechanism of working memory and the constitutive interpersonality of intelligence—is obscured. So long as there is a

viable nonreductive account of intelligence on the table, reduction carries no explanatory benefits and falls into at least two significant explanatory pitfalls.

And there are several viable nonreductive accounts on the table. For instance, rather than measuring a cognitive mechanism itself, perhaps *g* measures an effect of the interactions of several mechanisms. As several researchers have argued, there is good reason to believe that the positive manifold is "an emergent property of anatomically distinct cognitive systems, each of which has its own capacity" (Hampshire et al. 2012: 1225). At its extreme, this approach leads to the conclusion that "*g* is 'not a thing' but instead is a summary statistic" and thus that "the search for the neural basis of *g* is meaningless" (Conway & Kovacs 2018: 59). If viable, this approach would avoid both pitfalls of reducing intelligence to working memory: it wouldn't exclude features of the positive manifold on an *ad hoc* basis, and it would have the flexibility to countenance the constitutively comparative nature of the positive manifold. (After all, some emergent properties—like the property being taller than somebody else—emerge only in the light of a relation that undergirds comparisons. The target of a summary statistic is a perfect candidate for just such a constitutively comparative emergent property.)

3.3. Mutualism

In that spirit, van der Maas and colleagues have vigorously argued that the intercorrelations between individuals' IQ test scores can be explained by reference to the dynamic interplay of specialized cognitive mechanisms.

Van der Maas et al. (2006) analogize *g* to the results of predator-prey dynamics in ecology. According to the Lottka-Volterra model (Weisberg 2013), high correlations between

predator and prey populations needn't be caused by a single underlying factor (a shared food source, say) which bolsters both populations. Instead, the correlation can be caused—and in nature actually is often caused—by dynamic interactions between the two populations. The size of the prey population increases when the size of the predator population is small (because breeding outpaces being eaten), and decreases when the predator population is large (because being eaten outpaces breeding). At the same time, the predator population grows when the prey population is large (because eating causes breeding), and decreases when the prey population is small (because there isn't enough food to go around). These dynamics ensure that a strong correlation between the size of the populations emerges over time, without requiring any underlying factor to affect both populations.

Analogously, van der Maas and colleagues have demonstrated that high correlations between the performance of distinct cognitive mechanisms, which each undergird performance on some IQ subtest or other, needn't be caused by a particular underlying factor which fuels each performance. Instead, the correlations are plausibly caused by dynamic interactions between the distinct cognitive mechanisms. Research in cognitive psychology reveals that such dynamic relationships between cognitive processes exist. Short-term memory improves the development of cognitive strategies, and cognitive strategies improve the efficiency of short-term memory (Siegler & Alibali 2005). Language production and reasoning are similarly mutually beneficial: if you can think through it, then you can put it into words better, and if you can put it into words better, then that helps you think through it better (Fisher et al. 1994). And so on. These sorts of dynamic interactions between distinct cognitive mechanisms generate

positive feedback loops, ensuring that strong correlations emerge over time between how well mechanisms function across the cognitive system.

g is an explanandum, not the explanans, of the mutualistic functioning of cognitive mechanisms. If theorists force *g* into the role of explanans, then they'll find that it is, at best, a low-fidelity model of that functioning: it idealizes away from all of the independently interesting, messy and complex mechanistic details. Van der Maas and colleagues (2014) go on to infer that *g* is of *theoretical* interest only as something to be explained; it is a *predictively* powerful construct, but it doesn't itself do any interesting explanatory work.

4. *g* as bridge model

I think this last inference is mistaken. On my view, *g* does interesting explanatory work, *not* as a model of mechanisms, but as a *bridge model* that illuminates the relationship between folk psychological intelligence and the functioning of cognitive systems.

On Weisberg's (2013) influential account, models are (concrete, mathematical, or computational) structures plus construals—scientists' interpretations of those structures as descriptions of target systems. Bridge models are structures that scientists construe as describing the relationship between two or more target systems. Bridge models are particularly useful as aids to explanations of the relationships between two different levels (or otherwise incommensurate varieties) of scientific explanation. Most explanatorily powerful models idealize away many irrelevant features of their target systems. In the case of bridge models, this means ignoring many (if not all) of the features of each of the target phenomena that aren't directly related to the other target phenomenon.

My positive proposal is that the same idealizations and abstractions that render *g* a low-fidelity model of both folk psychological intelligence and its cognitive underpinnings also render it a high-fidelity bridge model. By distilling the common core of IQ-test-taking-ability, *g* idealizes away all of the details of cognitive functioning except the fact that cognitive systems produce a positive manifold. At the same time, *g* also idealizes away the aspects (indeed, whole varieties) of folk psychological intelligence that aren't tracked by performance on IQ subtests. Nevertheless, under the proper respective construals, *g* serves as a low-fidelity model of each of these phenomena. In so doing, it doesn't allow researchers to get a very firm grasp on either the folk psychology or the cognitive psychology of intelligence. But, properly construed, it could allow theorists to get a firmer grasp on the relationship between these two varieties of psychological explanation. In Sellarsian jargon: *g*, construed as a bridge model, can help fuse the manifest and scientific images of intelligence into one synoptic vision.

As construed by van der Maas, *g* doesn't provide a mechanistic explanation, but it does capture the fact that cognitive mechanisms dynamically work together to form a general substrate for the constitutively comparative problem-solving capacities that constitute the relevant variety of folk psychological intelligence. Taken from the other direction, *g* is, at best, a low-fidelity model of folk psychological intelligence: it idealizes away from the multifarious cross-cultural differences between folks' conceptions of intelligence, and from many of the messy and complex details within conceptions. Nevertheless, *g* is a high-fidelity model of those aspects of folk psychological intelligence that are realized by the mutualistic network of cognitive mechanisms that subserves IQ-test-taking-ability. When properly construed as a bridge model, *g* thereby helps reveal why and how one variety of lay intelligence attribution is

genuinely powerfully predictive (and in some senses explanatory) of human behavior. An idealization of the attributed suite of constitutively comparative problem-solving capacities maps onto a predictively fecund idealization of the dynamic interactions between cognitive mechanisms.

By the same token, treating *g* as a bridge model is explanatory of its own extremely high correlation with certain measures of success in life. *g* isn't a great measure of any particular aspect of cognitive functioning. Nor is it a great measure of any particular folk conception of intelligence. But it does help researchers zero in on those aspects of cognitive functioning—the relevant mechanisms and their interactions—that undergird core features of some culturally salient folk conceptions of intelligence. In other words, it is a great measure of the features of cognitive functioning that many people value when they value intelligence—and thus of the aspects of cognitive functioning that lead to certain kinds of success in a society partly structured by people's values.

Researchers make a mistake when they infer that *g* must be a great measure of cognitive functioning, since it is so predictive of success. On the contrary, we should expect *g* qua bridge model to correlate with success better than any great direct measure of cognitive functioning. After all, most folks (and their social institutions) don't care a wit about rewarding cognitive functioning per se—they care about rewarding those people whose cognitive functioning has put them in a position to accomplish valued goals. At the same time, we should also expect *g* qua bridge model to correlate with success better than any great direct measure of intelligence as it emerges in relation to any given folk conception, since it zeroes in on those aspects of folk

psychological intelligence that are actually undergirded by more or less efficient and effective cognitive functioning.

I'll conclude by drawing a concrete philosophical lesson. Psychofunctionalists have often argued that belief attributions must literally describe cognitive functioning, since they are predictively fecund (Fodor 1987; Quilty-Dunn & Mandelbaum 2018). There is something to this thought: folk psychological beliefs must be undergirded by reliable patterns of cognitive functioning. Nevertheless, *g*, as bridge model, clearly highlights how intelligence attribution is predictively fecund without literally describing cognitive functioning. Likewise, the predictive fecundity of belief attribution at most shows that, if we were to construct the relevant bridge model, we'd find a relationship between some aspects of folk psychological belief and some cognitive underpinnings that are responsible for behaviors that can be predicted via belief attribution. It can't show that folk psychological belief is reducible to those cognitive underpinnings: intelligence attribution is similarly predictively powerful despite being irreducible. Of course, this doesn't show that psychofunctionalism about belief is false. Some reductions of folk psychological phenomena to cognitive phenomena are well-founded. But I have argued that, intrapersonally speaking, human cognitive architectures don't feature anything well-labeled 'intelligence'. It is still an open question, which won't be settled by appeals to the predictive power of folk psychology, whether they feature anything well-labeled 'beliefs'.

References

Block, N. & Dworkin, G. (1974a). "IQ, Heritability and Inequality, Part 1. *Philosophy & Public Affairs* Vol. 3, No. 4, 331-409.

Borsboom, D., Kievit, R., Cervone, D., & Hood, S. (2009). The two disciplines of scientific psychology. In J. Valsiner, P. Molenaar, M. Lyra, & N. Chaudhary (Eds.), *Dynamic Process Methodology in the Social and Developmental Sciences*. Springer.

Carruthers, P. (2015). *The Centered Mind*. OUP.

Conway, A. & Kovacs, K. (2018). The Nature of the General Factor of Intelligence. In Robert Sternberg, ed. *The Nature of Human Intelligence*. CUP, 49–63.

Curry, D. S. (forthcoming). Street smarts. *Synthese*.

Deary, I, Penke, L., & Johnson, W. (2010). The Neuroscience of Human Intelligence Differences. *Nature Reviews: Neuroscience* 11: 201–211.

Dennett, D. (1991). Real Patterns. *The Journal of Philosophy*, Vol 88, No. 1, 27–51.

Dweck, C. & Bempechat, J. (1983). Children's Theories of Intelligence. In Paris, Olson, & Stevenson (Eds.), *Learning and Motivation in the Classroom*. Taylor & Francis.

Fisher, C., Hall, D., Rakowitz, S., & Gleitman, L. (1994). Syntactic and conceptual constraints on vocabulary growth. *Lingua* 92: 333–375.

Fodor, J. (1987). *Psychosemantics*. MIT Press.

Gottfredson, L. (1997). Why *g* Matters. *Intelligence* 24(1), 79–132.

Haier, R. (2017) *The Neuroscience of Intelligence*. CUP.

Hampshire, A., Highfield, R., Parin, B., & Owen, A. (2012). Fractionating Human Intelligence. *Neuron* 76, 6, 20, 1225–1237

Henrich, J., S. Heine & A. Norenzayan. (2010). The Weirdest People in the World? *Behavioral and Brain Sciences* 33, 61–135

Jaušovec, N. & Jaušovec, K. (2012). Working memory training. *Brain and Cognition* 79, 96–106.

Jensen, A. (2006). *Clocking the mind*. Elsevier.

Jung, R. & Haier, R. (2007). The Parieto-Frontal Integration Theory (P-FIT) of Intelligence. *Behav Brain Sci* 30(2): 135–187.

Mackintosh, N. (2011). *IQ and Human Intelligence*. Second Edition. OUP.

Quilty-Dunn, J. & Mandelbaum, E. (2018). Against dispositionalism. *Philosophical Studies* 175(9): 2353–2372.

Ryle, G. (1945). Knowing how and knowing that. *Proceedings of the Aristotelian Society* 46: 1–16.

Shipstead, Z. & Engle, R. (2018). Mechanisms of working memory capacity and fluid intelligence and their common dependence on executive attention. In Robert Sternberg, ed. *The Nature of Human Intelligence*. CUP, 287–307.

Siegler, R., & Alibali, M. (2005). *Children's thinking* (4th ed.). Prentice Hall.

Sternberg, R. & Grigorenko, E. (2004). Intelligence and culture. *Phil. Trans. R. Soc. Lond. B.* 359, 1427–1434.

Van der Maas, H., Dolan, C., Grasman, R. Wicherts, J., Huizenga, H., Raijmakers, M. (2006). A dynamical model of general intelligence. *Psychol. Rev*. 113, 842–861.

Van der Maas, H., Kan, K. & Borsboom, D. (2014). Intelligence is What the Intelligence Test Measures. Seriously. *J. Intell*. 2(1), 12–15.

Weisberg, M. (2013). *Simulation and Similarity*. OUP.

Implementation as Resemblance[1]

Abstract

This paper advertises a new account of computational implementation. According to the resemblance account, implementation is a matter of resembling a computational architecture. The resemblance account departs from previous theories by denying that computational architectures are exhausted by their formal, mathematical features. Instead, they are taken to be permeated with causality, spatiotemporality, and other non-mathematical features. I argue that this approach comports well with computer scientific practice, and offers a novel response to so-called triviality arguments.

1. Theories of Implementation

Theories of physical computation address two questions:

Q1. What distinguishes physical systems that compute from those that don't?

Q2. Among physical computing systems, what distinguishes those that compute the same thing from those that don't?

(1) concerns the difference between laptops and calculators on the one hand, and rocks and tables on the other. (2), by contrast, concerns the distinction between one laptop computing dot products and another computing Fourier transforms. An adequate account of physical computation should answer both (Sprevak 2019).

Different answers to Q1 and Q2 are possible. I shall be concerned with *implementationist* theories, which hold that a physical system computes if it implements some computational system, or 'computation', for short.[2] Thus:

[1] Draft of March 5, 2020. Word count: 4947.

[2] What of alternatives to implementationism? On one reading, Piccinini's (2015) mechanistic account answers Q1 and Q2 by direct appeal to the notion of a computing mechanism. So construed, the implementation relation plays no part in the mechanistic account. This sort of approach is worth exploring, but is beyond the present scope.

A1. A physical system computes just in case it implements some computation.

A2. What a physical system computes is determined by the computation it implements.

These answers are schematic, however. They say little about what computations are, and what implementation amounts to. Different accounts of implementation emerge from different specifications of these details.

This paper introduces a new account implementation, which I call the *resemblance account*. I sketch the account in Sections 2 - 4. Section 5 deals with some background metaphysical issues. Sections 6 argues that the resemblance offers an interesting new perspective on some old problems in the philosophy of physical computation. Finally, Section 7 deals with an objection, and Section 8 concludes.

A caveat before proceeding. My main aim is advertisement: to show that the resemblance account offers a novel approach to physical computation, worthy of further investigation. Regrettably, however, this means some issues won't receive the treatment they deserve. These issues must wait for another occasion, and I'll flag them as they arise.

2. The Resemblance Account

I propose to begin at the beginning. In his landmark 1936 paper, Turing offers the following description of an *a*-machine:

> The machine is supplied with a "tape " (the analogue of paper) running through it, and divided into sections (called "squares") each capable of bearing a "symbol" … the configuration [of the machine] determines the possible behaviour of the machine. In

> some of the configurations in which the scanned square is blank (i.e. bears no symbol) the machine writes down a new symbol on the scanned square: in other configurations it erases the scanned symbol. The machine may also change the square which is being scanned, but only by shifting it one place to right or left. (Turing 1936, 231)

As we know, Turing arrived at this conception of *a*-machines by carefully considering the activity of human workers proceeding effectively. The restriction that *a*-machines may only 'observe' one symbol at a time, for instance, is justified on the grounds that human workers can only distinguish between finitely many different primitive symbol types. (For, if not, then we could distinguish between (tokens of) types which differ to an arbitrarily small degree. But our perceptual apparatus is not nearly as sophisticated as this. See Turing (1936, 249 - 252) and Sieg (2009) for discussion.)

The importance of Turing's insight is not hard to appreciate. By linking the characterization of an *a*-machine directly to the activities of actual human workers, Turing's analysis sheds light on the computational capacities and limitations of humans working effectively. Very roughly, the computational power of *a*-machines bears on the computational power of effective human workers because the former resemble the latter in certain important respects: both have certain 'perceptual' limitations, both follow only finitely many instructions one at a time, and so on. Indeed, alternative analyses, such as λ-definability or Herbrand-Godel general recursivity, were unsatisfactory because they fail to adequately illuminate the basic activities of a human working effectively.

I mention all of this because it seems to me that Turing's analysis contains the essentials of the resemblance account. *A*-machines bear on the computational powers of human workers because, and to the extent that, the former resemble the latter in certain respects. The resemblance account generalizes and precisifies this insight. On the resemblance account, physical computation is a matter of resembling a computational architecture:

> **The Resemblance Account.** A physical system computes just in case, and to the extent that, it resembles a computational architecture.

In the following two sections I flesh out the notion of a computational architecture, and explain the notion of resemblance at play. But it should be noted that while I talk of *the* resemblance account, really I am scouting a family of views. Different ways of filling in the sketch I give deliver different particular resemblance accounts. I will mention the major choice-points as they arise.

3. Computational Architectures

Turing's *a*-machines are an example of what I'll call a *computational architecture*. To a first approximation, computational architectures are ‘blueprints’ for physical computing devices. Blueprints 'specify' which features a system must have in order to count as a computing device of a particular sort (I return to the question of what ‘blueprints’ are, and what 'specification' amounts to, in Section 5). Turing’s description, for instance, constitutes a blueprint which specifies the features a physical system must have in order to ‘counts as’ an *a*-machine.

Accordingly, a physical system performs *a*-machine computations only if it has these features too. Anything lacking these features doesn't count as an *a*-machine, hence *a fortiori* doesn't perform *a*-machine computations either.

But what are these features? Some concern the physical or mechanical features of the device. For instance, Turing's characterization requires that *a*-machines *scan* the tape, *write* symbols, and *shift* left or right, that they have a *read/write head* and a *tape* divided into squares, and that latter machine states be *determined* by earlier states. Other features are more abstract, and concern the patterns or regularities the machine or its components exhibit. Others still concern what states of the device represent. The symbols on the tape refer to natural numbers, for instance, and a machine as a whole may be taken to represent, in some sense, the function it computes. The upshot of all of this is that a physical must exhibit these sorts of features if it is to 'count as' an *a*-machine, or if it is to perform *a*-machine computations.

However, while Turing's characterization is illustrative, it is not representative of contemporary computer design. For a state-of-the-art understanding of computer architectures (in the present sense), we should look to work on computer architecture and engineering. These disciplines truck in highly specific descriptions of computational architectures. For instance, for a physical system to count as a MIPS (Microprocessor without Interlocked Pipelined Stages) microarchitecture, it must exhibit a highly specific set of features.[3] Some of these features are described explicitly in the microarchitecture description, for instance that the system have a datapath with a certain pipelining scheme, certain components for sign extension operations, and so on. Others are left tacit, such as the requirement that the system be cast in a silicon wafer, that

[3] See Harris and Harris (2013).

it have a certain clock rate, and so on. As in the *a*-machine case, nothing counts as a MIPS microarchitecture unless it has these features.

Stepping back, it seems to me that three kinds of features are commonly cited in the descriptions offered by computer architects and engineers. *Physico-mechanical features* concern the physical and micro-physical structure of a system: its components and their relationships, interactions, and composition. Other features concern the patterns or regularities exhibited by various states and components, and I'll call these features *syntactic*. Finally, *semantic* or *representational* features concern what the states or processes of the device represent. However, I don't take this list to exhaust the features that may be specified by a computational architecture. Indeed, it would be a mistake to try to specify such a list once-and-for-all. Instead, we should regard this list as open to addition or amendment, as computer engineers and architects devise new sorts of computing systems with new and different sorts of features.

Computational architectures can be more or less fine-grained. Some omit irrelevant details, as when they indicate that one must build a column that supports 500kg, but does not say whether it must be made of wood or stone. Others are more exacting, and demand that a five meter tall fluted marble column be put here with such-and-such capital ornaments. Similarly, in the computational case we might be told that the device is to have a read/write head, but not told what it is made of. Other times we are told that the device must be made of silicon, have 500MB of L3 cache, have four cores, run at 2.4GHz, and so forth.

However, the question of which specific physico-mechanical, syntactic, or semantic properties are required for computation is best left to computer scientists and engineers. It is a

highly non-trivial task designing a computing system, involving a tremendous amount of epistemic labour.[4] Consequently, I doubt that philosophers have much to contribute on this front.

4. Resemblance

The next task is to say what it is for a physical system to 'resemble' a computational architecture. This marks a choice point in the theory. Some philosophers might be content to rest with an intuitive, or commonsensical, notion of resemblance. This is fine as far as it goes, but there are benefits to working with a more precise account, and here I will offer one.

To a very rough first approximation, the idea pursued here is that a physical system resembles a computational architecture to the extent that it (a) has features 'specified' by the architecture, and (b) lacks features *not* 'specified' by the architecture. A physical system resembles an *a*-machine, for instance, to the extent that it has a read/write head, a control unit, tape, and so on. (What it is for an architecture to 'specify' a feature depends to some extent on one's background metaphysics; see Section 5 for more.)

Recent work on similarity can be used to make this precise.[5] On this theory, resemblance is always determined relative to a distinguished class of features *F*, called the feature set. If *C* is a computational architecture and *P* is a physical system, we'll let $F_C \subseteq F$ be the features specified by *C* and $F_P \subseteq F$ be the set of features of *P*. Then we can say that *P* resembles *C*, with respect to *F*, to degree *n*, just in case

[4] To get a sense of the complexity involved, note that a modest contemporary microprocessor houses approximately 4.5 billion transistors, and executes upwards of 140 instructions in parallel.

[5] The account presented here follows Weisberg's 'weighted feature-matching account' (2012). For recent criticisms and elaborations, see Parker (2015) and Fang (2017). Some philosophers distinguish between resemblance and similarity, but here I use the terms interchangeably.

$$|F_C \cap F_P| - |F_C - F_P| - |F_P - F_C| = n \quad (1)$$

For convenience, I'll write *Res(F, C, P) = n*. Here $F_C \cap F_P$ are the features shared by *C* and *P*. $F_C - F_P$ are the features specified by *C* which *P* lacks. And $F_P - F_C$ are the features had by *C* not specified by A. This equation, in effect, measures the extent to which *P* 'fits' *C*'s specification.

This account treats resemblance as a graded notion. I think this is the most basic notion of resemblance, and we can use it to define other notions as the need arises. For instance, we can say that *P perfectly resembles C* just in case *Res(F, C, P) = n* and $|F_C \cap F_P| = n$. Similarly, we can say that *C* and *P resemble each other simpliciter* just in case *Res(F, C, P)* $\geq m$, for some predetermined 'cutoff' degree of resemblance *m*. For present purposes I think it is enough to work with the basic notion, but I am open to the possibility that a more refined notion is appropriate for thinking about implementation.

Given this account of resemblance, implementation is in the first instance a matter of degree, so that a physical system implements a given computational architecture to a greater or lesser degree. Some philosophers might be uncomfortable with this result, preferring an absolute notion of implementation instead. But as I just mentioned, given a graded notion of implementation we can use it to define an absolute notion if we want. Moreover, there is some independent reason for working with a graded notion. Computer scientists often talk about different physical systems being better or worse implementations of a given architecture. This practice is naturally understood as relying on a graded notion of implementation, and it's not clear how to capture this talk, without distortion, with an absolute notion.

One further issue deserves comment. Exactly what degree of resemblance is required for implementation? On the one hand, perfect resemblance seems too exacting: it seems useful to allow that a physical system may implement an architecture even when they don't perfectly resemble each other. On the other, too low a degree threatens to trivialize the notion of physical computation: there are plausibly some simple physico-mechanical properties shared by paradigmatically non-computing systems and any computational architecture. With this complication flagged, I will simply say that implementation requires a 'sufficiently high' degree of resemblance, noting that this is just a placeholder for what will undoubtedly be a complicated theory of just what a 'sufficiently high' degree amounts to.

5. Interlude: Matters of Metaphysics

So far I've glossed computational architectures as 'blueprints', and I've said that blueprints 'specify' features. But what does all this mean? Presumably we don't wish to add 'blueprints' as a new fundamental to our ontology, so we'd better find a way to cash them out in more familiar terms.

Perhaps unsurprisingly, this marks another choice point for the theory. There are different ways of cashing out the notion of a blueprint, according to different tastes in background metaphysics. Here I'll mention two, but I don't take these to exhaust the alternatives. In fact, I think the resemblance account can be developed in a way that accommodates a wide variety of views in metaphysics, and I take this to be a virtue of the account.

One option takes computational architectures to be highly specific universals. In this case, implementation boils down to instantiation, and a computational architecture 'specifies' a

class of features in something like the way that conjunctive universal 'specifies' its conjuncts, for instance by having them as parts. So construed, Turing's description of *a*-machines is a description of a universal, the instantiation of which is a matter of having a read/write head, a tape, and so forth. A physical system will instantiate a computational architecture more or less, to the degree that it instantiates the conjuncts that compose the architecture in question. In this case, *Res(F, C, P)* is a measure of the degree of instantiation of a given universal.

Philosophers who balk at universals will prefer a more deflationary approach. Here too there are many options available. One treats computational architectures as abstract particulars, perhaps in something like the way that some scientific models are said to be abstract particulars.[6] On this account, computational architectures are taken to literally *have* certain physico-mechanical, syntactic, or semantic features. In this case, resemblance amounts to property sharing, so that *Res(F, C, P)* is a measure of the degree to which *P* has features also had by *C*. And other deflationary approaches are possible too. For instance, we might take computational architectures to be linguistic entities -- descriptions, say -- so that implementation boils down to some sort of semantic relation, such as accurate description. Degree of resemblance then amounts to how well a given computational architecture describes a given physical system.

At any rate, which way you go turns, to a large extent, on your ontological tastes. I take it to be a virtue of the resemblance account that it can be developed in a way that satisfies a wide variety of ontological palates.

[6] See, e.g., Giere (1988, Chapter 3). A related approach, although arguably more deflationary, is found in Godfrey-Smith (2009), who suggests that models are fictional entities. Copeland and Shagrir mention, but do not endorse, a view of computation in this vicinity too. As they explain, their view "recognizes an ontological level lying between the realization (or physical-device) level and the level of pure-mathematical ontology ... At this level are to be found notional or idealized machines that are rich with spatio-temporality and causality" (2011, 234).

6. Triviality Arguments

So far I've motivated the resemblance account by noting that it matches the thought and talk of computer scientists. Computer scientists routinely describe computational architectures, such as Turing machines, in spatio-temporal terms. I take it to be a point in favour of the resemblance account that it reflects this practice. But the resemblance account is also attractive on theoretical grounds. In particular, the resemblance account offers a novel response to so-called triviality arguments. This section sketches that response.

It is instructive to recall how triviality worries emerge for standard theories of implementation. The most popular theory holds that implementation is a relation between a physical system, on the one hand, and an abstract, mathematical computation on the other. In the simplest case a computation is a finite automaton, composed of a finite set of states plus a transition function. More complicated computations may include inputs, outputs, or states with internal combinatorial structure. These details aside, however, the characteristic feature of the received view is that computations are exhaustively characterized by their formal, mathematical structure. Following Rescorla (2014), I will call this *structuralism about implementation*.[7]

Structuralism holds that a physical system implements a computation if its state transitions 'mirror' the state transitions of some formal computation. 'Mirroring' is typically taken to be a structure-preserving map, or isomorphism, between physical and formal states. Thus, according to a simple structuralist view, if C = (S, T) is a computation with states S = {S_1, S_2, …, S_n} and transition function T: S → S, we have:

[7] Chalmers (1996) is representative of this approach. Other proponents include Millhouse (2017), Schweizer (2019), Sprevak (2010), and Scheutz (2001), among many others.

> **Structuralism.** Physical system *P* implements computation *C* just in case there exists a mapping *f* from states of P to S such that: if P is in a state P_i for which $f(P_i) = S_k$, and $T(S_k) = S_m$, then P goes into state P_j for which $f(P_j) = S_m$.

Notoriously, however, mappings are cheap and computations are abundant. For nearly every physical system P and every computation C, there is a structure-preserving map between P and C, in which case nearly every physical system implements every computation, according to (SF). This is the core of the triviality worry.[8]

In light of triviality worries, few philosophers endorse structuralism in this unalloyed form.[9] Various additional constraints are added to (SF) in an effort to avoid triviality. Some common requirements are that implementing systems satisfy counterfactual conditionals (Block, 1995; Copeland, 1996); that they exhibit appropriate causal structure (Chalmers, 1996; Scheutz, 2001); that distinct physical states be mapped to distinct formal states (Chalmers, 1996; Godfrey-Smith, 2009); that only appropriately `natural' or 'simple' physical states feature in the mapping (Scheutz, 2001; Godfrey-Smith, 2009; Millhouse, 2017); that the physical states have representational properties (Shagrir, 2001, 2018; Sprevak, 2010); or that the physical states be states of functional mechanisms (Piccinini, 2012, 2015). And others are surely possible.

All of these tactics are a kind of 'bottom up' response to triviality. They attempt to cut down the class of implemented computations by constraining which physical systems figure in the preimage of the implementation mapping. Only those with the appropriate counterfactual,

[8] See Sprevak (2019) for more.
[9] Schweizer (2019) is a recent exception.

representational, etc. features get in. A 'top down' response, by contrast, enriches the account of the systems in the image of the mapping. To the extent that formalists take computations to be formal, mathematical objects, they *must* take bottom up approaches to triviality. The few 'top down' approaches they can take typically complicate the structure of computations while preserving their overall formal, mathematical character.

Concerning triviality, the resemblance account employs a 'top down' response. The account denies that computational architectures are exhausted by their formal, mathematical features. Instead, they may be replete with physico-mechanical, representational, etc. features. Since implementation is a matter of resembling a computational architecture in these respects, and since most physical systems *don't* have the right arrangement of features, most physical systems won't implement many, or even any, computational architectures. For instance, most physical systems don't have a read/write head, an indefinitely extensible tape, etc., so most systems won't implement *a*-machines. Similarly, most physical systems aren't composed of a silicon board, have a certain pipelining scheme, and so on, so most physical systems don't implement a MIPS microarchitecture.[10]

There is some reason to be dissatisfied with 'bottom up' approaches; here I'll mention just one. The worry is that structuralism cannot adequately explain *why* physical computation necessarily involves causal, semantic, etc. features.[11] From the structuralist's perspective, computation is fundamentally a mathematical phenomenon, captured by pure mathematical computing systems such as Turing machines (construed set-theoretically), DFAs, and the like.

[10] However, attentive readers will note that this response turns, to some extent, on the specific choice of features in question. There subtleties must await another occasion.
[11] Another is that structuralism fails to capture the implementation conditions of many computational models; see Rescorla (2014).

Methodologically, the strategy is to *start* with a prior mathematical notion of computation, and define a notion of physical computation in terms of it.[12] But given this outlook, the requirement that physical computation essentially involve causal, semantic, etc. features looks less like a discovery about the nature of physical computation, and more like an *ad hoc* maneuver designed to save the theory from triviality. If computation is fundamentally a mathematical phenomenon, as the structuralist holds, and if that account of computation leads to trivialization (as it appears to), then what could possibly explain why physical computation *must* be causal, semantic, or whatever?

This explanatory problem doesn't arise for the resemblance account. Because the resemblance theorist denies that computation is essentially mathematical, there is no *additional* task of explaining why physical computation must involve causal-mechanical, representational, etc. features. Recall Turing's characterization of *a*-machines. On that characterization, it is constitutive of *a*-machines that they have a read/write head, that later states be determined by earlier states, and so on. From this perspective, *what it is* to be an *a*-machine is just to have these features. But since this characterization comes with causal-mechanical features 'built in', so to speak, it is straightforward to explain why a physical system must have these features in order to carry out *a*-machine computations. The reason is simply that *a*-machine computations *just are* a certain kind of causal-mechanical (etc.) process. From this perspective, a causal-mechanical requirement isn't an *ad hoc* maneuver designed to save the theory from triviality, but instead reflects a basic fact about the nature of *a*-machine computations, namely, that they are a kind of causal-mechanical process.

[12] See Chalmers (1994, 341-342) for an especially clear statement of this approach.

7. Medium Independence

The resemblance account responds to triviality by enriching the character of implemented computations. But this maneuver may seem to cut against the idea, widely endorsed, that computations are 'medium independent' (Piccinini 2015, ch. 7). This section explains how a suitable stand-in for medium independence can be developed within the resemblance framework.

To a first approximation, a property or process is medium independent if it can be realized in different physical media. *Cooking lentils* is not medium independent, because it can be realized in only quite specific physical media; *powering a drivetrain* is, because it can be accomplished by otherwise quite different physical systems (internal combustion engines, electric motors, etc.).

Medium independence is closely related to multiple realizability. A property or process is multiply realizable, roughly, if it can be realized by different kinds of physical systems. Medium independence entails multiple realizability: if a property or process is medium independent, then it can be realized in different physical media. Note, however, that the converse fails. *Being a corkscrew* is multiply realizable, since many different corkscrew designs might do the trick, but not medium independent. *Being a corkscrew* is a matter of interacting with a specific physical medium, namely cork.

It seems undeniable that computations are medium independent, hence multiply realizable. As Ned Block once pointed out, for instance, an AND-gate might be realized either by transistors, or by mice, string, and cheese (Block 1995). Moreover, the literature on unconventional computation is replete apparent cases in which the same computation (e.g., a sorting task) is performed by wildly different physical systems. But it's not clear that the

resemblance account can capture this apparent datum. The trouble is that there appears to be no single computational architecture, in the above sense, common to the wide variety of computing systems hypothesized by computer scientists. There is no architecture, for instance, which both silicon and murine AND gates resemble.[13]

What should the resemblance theorist make of this? The solution, I think, becomes clear once we reflect on the role played by medium independence (multiple realizability) in computational theorizing. In general we require a way to describe physical systems that abstracts away from (some of) their physical details. So abstracted, we can consider whether, e.g., two systems compute the same logical function despite physical dissimilarities. Now, ordinarily this role is played by the alleged medium independence (multiple realizability) of computations. But if the resemblance account can supply a way to abstract from physical details, it can furnish a way to describe physical systems at the desired level of abstraction. And this, I submit, is just what is needed for computational theorizing. The rest of this section explains how this might go.

To begin, while above resemblance is characterized in terms of a single class of features, we can also define a notion of resemblance that discriminates between different classes of features. If $F_1, F_2, \ldots, F_m$ are classes of features, then we can say that P resembles C with respect to F_i $(1 \leq i \leq m)$ to degree n just in case $\sum_{i=1}^{m} Res(F_i, P, C) = n$. Moreover, by adding a coefficient to equation (1) we can discount (or boost) the contribution of a particular class of features to the overall resemblance score.[14] Doing so gives

[13] There is, I suppose, a *disjunctive* architecture composed of both silicon and murine components. But such a device is a metaphysician's contrivance, not a genuine deliverance of computer science.
[14] Cf. Weisberg (2012).

$$x(|F_A \cap F_P| - |F_A - F_P| - |F_P - F_A|) \quad (2)$$

I will write $Res(F, P, C) = n$ as shorthand. In general, then, if $x_i \in \mathbb{R}$ $(1 \leq i \leq m)$ are coefficients, the generalized resemblance score is given by $\sum_{i=1}^{m} Res(F_i, x_i, P, C) = n$.

By appropriately choosing coefficients we can define a notion of *pattern resemblance* between systems. For instance, if F_1, F_2, F_3 are classes of physico-mechanical, syntactic, and semantic features, respective, with corresponding coefficients x_1, x_2, x_3, then by setting $x_1 = x_3 = 0$ and $x_2 = 1$ we can say that P *pattern resembles* C to degree *n* just in case $\sum_{i=1}^{3} Res(F_i, x_i, P, C) = n$, that P *pattern resembles C* simpliciter just in case P pattern resembles C to a high enough degree, and so on.

How does all this help? Medium independence is naturally thought to concern what I've called 'syntactic' features. We say that silicon and mouse-and-string systems compute AND, when they do, because at a certain abstract level of description they exhibit the same patterns, regardless of their physical substrate. The resemblance account can accommodate this fact by noting that different AND gates pattern resemble each other. Thus the resemblance account can furnish a level of description appropriate for this part of computational theorizing, without abandoning the insight that inclusion of specific physico-mechanical features is central to avoiding triviality.

8. Summary

The resemblance account holds that a physical system computes to the extent that it resembles a computational architecture. This view is grounded in Turing's influential approach to thinking about computational architectures, an approach which persists in computer science today. The view is also motivated on theoretical grounds: it offers a novel and natural response to triviality arguments about computational implementation. While I haven't here attempted an exhaustive assessment of the resemblance account, the considerations surveyed here are promising. For this reason I take the resemblance account to be worthy of further investigation.

References

Block, N. (1995). The Mind as the Software of the Brain. In *An Invitation to Cognitive Science, 2nd ed, Vol 3*, Ed. Osherson et al. MIT Press, 377-425.

Chalmers, D. (1994). On implementing a computation. *Minds and Machines*, *4(4)*, 391-402.

Chalmers, D. (1996). Does a Rock Implement Every Finite-State Automaton? *Synthese, 108(3)*, 309-333.

Copeland, B. J. (1996). What is Computation? *Synthese 108*, 335-359.

Copeland, B. J. & Shagrir, O. Do Accelerating Turing Machines Compute the Uncomputable? *Minds and Machines, 21(2)*, 221-239.

Fang, W. (2017). Holistic modeling: an objection to Weisberg's weighted feature-matching account. *Synthese 194*, 1743-1764.

Giere, R. N. (1988). *Explaining Science: A Cognitive Approach*. University of Chicago Press.

Godfrey-Smith, P. (2007). Models and Fictions in Science. *Philosophical Studies 143(1)*, 101-116.

Godfrey-Smith, P. (2009). Triviality arguments against functionalism. *Philosophical Studies, 145(2)*, 273-295.

Harris, D. M., & Harris, S. L. (2013). *Digital Design and Computer Architecture* (2nd ed.). Morgan Kaufmann.

Millhouse, T. (2017). A Simplicity Criterion for Physical Computation. *The British Journal for the Philosophy of Science, online first*.

Parker, W. (2015). Getting (even more) serious about similarity. *Biology & Philosophy, 30*, 267-276.

Piccinini, G. (2008). Computation without Representation. *Philosophical Studies, 137(2)*, 205-241.

Piccinini, G. (2012). Computationalism. In *The Oxford Handbook of Philosophy of Cognitive Science*, Ed. Samuels et al. Oxford University Press, 222-249.

Piccinini, G. (2015) *Physical Computation: A Mechanistic Account*. Oxford University Press.

Rescorla, M. (2014). A theory of computational implementation. *Synthese*, *191*(6), 1277–1307.

Scheutz, M. (2001). Computational versus Causal Complexity. *Minds and Machines, 11(4)*, 543-566.

Schweizer, P. (2019). Triviality Arguments Reconsidered. *Minds and Machines*, DOI: 10.1007/s11023-019-09501-x.

Shagrir, O. (2001). Content, Computation and Externalism. *Mind 110(438)*, 369-400.

Shagrir, O. (2018). In defense of the semantic view of computation. *Synthese*. DOI: 10.1007/s11229-018-01921-z.

Sieg, W. (2009). On Computability. In: *Handbook of the Philosophy of Science, the Philosophy of Mathematics*, Ed. Irvine, A. Elsevier.

Sprevak, M. (2010). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science Part A, 41(3)*, 260-270.

Sprevak, M. (2019). Triviality arguments about computational implementation. In *Routledge Handbook of the Computational Mind*, ed. Colombo, M. Routledge, 175 - 191.

Turing, A, M. (1936). On Computable Numbers, with an Application to the *Entscheidungsproblem. Proceedings of the London Mathematical Society, 42(1)*, 230-265.

Weisberg, M. (2012). Getting Serious about Similarity. *Philosophy of Science*, *79*(5), 785–794.

DRAFT for PhilSci Archive: 7/30/2020 Mike Dacey

Anecdotal Experiments: evaluating evidence with few animals

Mike Dacey

Comparative psychology came into its own as a science of animal minds, so a standard story goes, when it abandoned anecdotes in favor of experimental methods. However, pragmatic constraints significantly limit the number of individual animals included in laboratories experiments. Studies are often published with sample sizes in the single digits, and sometimes samples of one animal. With such small samples, comparative psychology has arguably not actually moved on from its anecdotal roots. Replication failures in other branches of psychology have received substantial attention, but have only recently been addressed in comparative psychology, and have not received serious attention in the attending philosophical literature. I focus on the question of how to interpret findings from experiments with small samples, and whether they can be generalized to other members of the tested species. As a first step, I argue that we should view studies with extreme small sample sizes as *anecdotal experiments*, lying somewhere between traditional experiments and traditional anecdotes in evidential weight and generalizability.

1. Animal Anecdotes and the Founding of Comparative Psychology

Darwin's views on evolution suggest that continuity across species is the rule. Evolution occurs when small changes build up slowly over long periods of time, so we should expect to see cross-species continuity in most traits. Nowhere was this result more significant than when it came to the mind. The fiercely-held conventional wisdom at the time was that human minds were entirely unlike animal minds. To challenge this conventional wisdom, Darwin reports anecdotes about various clever and heroic animals. For instance:

> "I will give only one other instance of sympathetic and heroic conduct in a little American monkey. Several years ago a keeper at the Zoological Gardens, showed me some deep and scarcely healed wounds on the nape of his neck, inflicted on him while kneeling on the floor by a fierce baboon. The little American monkey, who was a warm friend of this keeper, lived in the same large compartment, and was dreadfully afraid of the big baboon. Nevertheless, as soon as he saw his friend the keeper in peril, he rushed to the rescue . . ." (1871 pg. 75)

This anecdotal approach continued in the work of George Romanes, Darwin's appointed successor on psychological topics. Describing similar animal heroism, Romanes says (also reporting the story secondhand) that a column of ants "rushed to the rescue" of an individual pinned with a rock, and "This

observation seems unequivocal as proving fellow-feeling and sympathy, so far as we can trace any analogy between the emotions of the higher animals and those of insects" (1888 pp. 48-49).

Near the turn of the 20th century, authors such as C. Lloyd Morgan (1894) and Edward Thorndike (1911) vocally disproved of the reliance on anecdotes. To be a science on firm founding, they felt, the field would need to shift to rigorous experimental methods. The resulting shift, so a common story goes, brought comparative psychology into its own as a rigorous science (e.g. Shettleworth 2012).

It is easy to see what is objectionable about the way Darwin and Romanes use anecdotes. They relay the stories secondhand without scrutiny, and leap to a heroic interpretation without considering other explanations. There is also a particular worry that work on animal minds will be systematically biased by the unconscious human tendency to anthropomorphize; to interpret animal actions in the same ways they would interpret human actions (e.g. Dacey 2017). Narrative anecdotes seem particularly ripe for such a bias. They often presume intentions behind the action (as when we describe a reach *for* an object, or a glance *towards* a person), and often elicit emotional reactions and bonds with characters that may threaten impartial scientific analysis.

To put it simply, rejecting anecdotes makes comparative psychology look more like other successful sciences (e.g. Thorndike 1911). Scientists across fields shun anecdotes. There are many reasons to do so. I attempt to summarize the key concerns about anecdotes below, listed to aid later discussion. These concerns overlap, and are not exhaustive:

1. Anecdotes can be cherry-picked to make a predetermined point.
2. We lack control over and knowledge of background conditions of anecdotes.
3. Anecdotes are narrative in structure, rather than providing analyzable data.
4. Anecdotes are non-repeatable (non-replicable), and so can't be confirmed independently.
5. Anecdotes don't support generalization.

Performing controlled experiments can alleviate these concerns. One cannot pick and choose which individual responses in any given experiment to report (though one can choose which experiments to report, as discussed below). A good experiment is defined by control over the variables that might influence behavior. Experiments produce evidence in the form of data, which is cold, dispassionate, and suited for statistical analysis. As a result, when done well, experiments are replicable (worries noted in section 3), and they can support generalization.

Summing up, anecdotes are usually opposed to experiments. A common foundation story for comparative psychology tells that it came into its own as a science when it chose experiments over anecdotes. However, it is not clear whether this foundation story holds up when we look at current practice.

2. Sample Sizes in Animal Labs

When running laboratory experiments on animals, practical constraints significantly restrict sample sizes. Animals must be kept and cared for, and labs can only afford and fit a certain number. Ethical concerns often dictate that the number of animals involved should be as low as possible.[1] Individual experiments usually require time-consuming training, so some subset of the overall groups is chosen.[2] There are also often basic tasks that an animal must successfully perform to even participate in the experiment, and those of the original group chosen who fail will be excluded. I take these to be challenges intrinsic to the subject of study, and do not intend to criticize the researchers who face them. Nonetheless, the implications are stark. Experiments frequently include samples of individual animals in the single digits, and sometimes only 1 or 2. Figure 1 shows the number of individual animals included in every individual experiment published in four top journals in the field in 2019. Out of 151 experiments in 90 papers, 50 experiments include data from 10 or fewer animals (nearly 1/3 of the total), and 98 include

[1] Both of these issues are especially difficult with primates, and even more so with chimpanzees, as in my example below.

[2] Additionally, having been trained on one task may influence later performance on other experiments, so sometimes animals are excluded so that they remain 'naïve' to the tasks at hand.

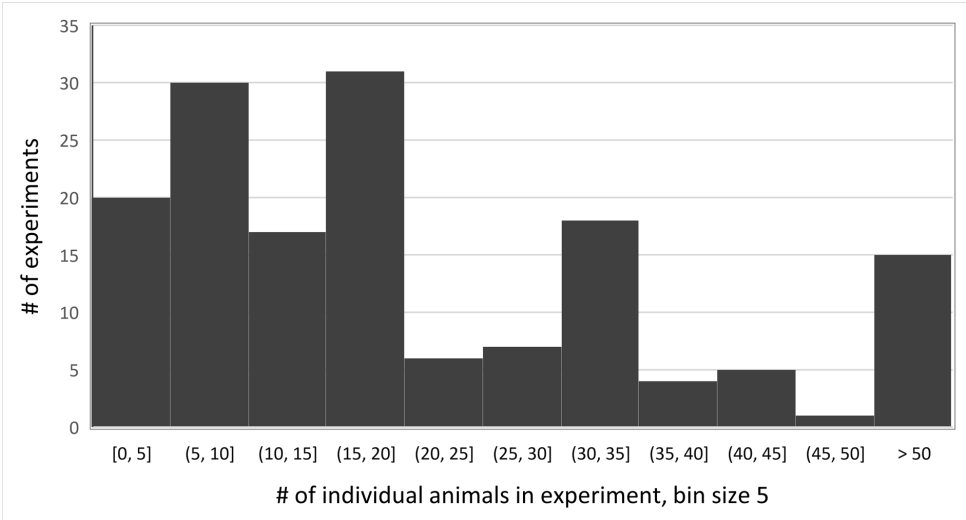


Fig. 1: A histogram of all experiments published in the journals *Animal Behavior and Cognition*, *Animal Cognition, Journal of Comparative Psychology,* and *Journal of Experimental Psychology: Animal Learning and Cognition* in the year 2019, sorted by number of individual animals in reported data. This includes 151 experiments described in 90 papers; any experiment that made an intervention, laboratory or field. 54 papers were excluded, as they did not present new behavioral data (29 papers), were unable to report the number of animals involved (seven papers), were purely observational field studies (six papers), or used only human subjects (12 papers).

data from fewer than 20 (nearly 2/3).[3] To put it bluntly, these sample sizes would be unacceptable in other branches of psychology.

As an illustrative example of the interpretive challenges raised by sample sizes like these, I will focus on Inoue & Matsuzawa's 2007 paper, "Working memory of numerals in chimpanzees." This paper compares human and chimpanzee performance on a short-term memory task. The authors state their conclusions unequivocally: "Our study shows that young chimpanzees have an extraordinary working memory capability for numerical recollection better than that of human adults" (pg. 1005). The paper has

[3] Thanks to Abraham Brownell for performing this analysis. This data is not meant to present a statistically rigorous picture of the field at large, but simply to provide a reasonably representative snapshot. This illustrates the issue to those unfamiliar with the norms of the field. These journals are among the top that focus on animal cognition, and were chosen in large part to limit potentially subjective inclusion criteria. However, they are not the only such journals, and animal cognition studies are often published in more generalist journals as well (for instance, the example discussed below was published in *Current Biology*). Several of these experiments also divided participants into different conditions, further limiting the number of individuals observed making specific responses, though we did not analyse these divisions.

been cited extensively, and in the media, this conclusion was accepted uncritically ("Chimps Exhibit Superior Memory, Outshining Humans," *New York Times* 12/4/2007).

The task was as follows. Participants (human and chimpanzee alike) sit in front of a computer screen. The computer quickly flashes several digits in random locations on the screen (all shown simultaneously). After a presentation of a few hundred milliseconds (650, 430, and 210 *ms* in different trials), each digit is masked with a small white square. Participants were asked to then tap each masking square in order of the digits previously at each location. The researchers measured both response times and accuracy. The task is meant to test the ability to rapidly store working memories for the visual scene (210 *ms* is too fast to saccade through the sequence).

Inoue and Matsuzawa begin the study with 6 chimpanzees (three mother-child pairs; there were 14 total on-site). While all six were able to learn the basic masking task, only four performed at the level of five numerals, which was the number used in the key test (Supplemental materials Table S1). So, the experiments include these four animals. The actual data presented, however, only compares one chimpanzee at a time against a human average (human n=9 in one experiment, n=12 in another). So for each actual comparison, chimpanzee n=1. In fact, the assertion that chimpanzees perform better than humans seems to be based on a single chimpanzee, Ayumu, the best chimpanzee performer (see figure 2). Based on the data presented in supplemental material (see figure 3), Ayumu matched the human average accuracy rate with 650 ms presentation times, but still had a lower accuracy rate than the majority of the individual humans.[4] So they key claim here seems to be based on a simple size of one.

Given this reliance on extremely small sample sizes, we must question whether the field has really moved on from its anecdotal roots. I suggest that performance of animals like Ayumu is just another kind of anecdote; it's a single animal (or very small number) displaying an interesting behavior. It can be hard

[4] All three chimpanzees shown did show faster response times than all humans (response time was measured as the latency before the first number was touched).

to know exactly what conclusions we can draw from a study like this. At the very least, though. findings like this cannot ground general claims like "chimpanzees outperform humans."[5]

This study is a particularly salient example, both in the sense that it reaches the limit case of n=1, and in its strong conclusion and broad uptake. But the core concerns here generalize, given the number of experiments published with extremely small sample sizes. To be clear, these restrictions result from practical issues intrinsic to the field. I do not criticize researchers for this, as I see no reasonable way around it (absent massive funding increases and means to address ethical concerns).

3. The replication crisis and comparative psychology

In recent years, other branches of psychology have instituted reforms to address prominent and repeated replication failures (Romero 2019). Despite the obvious worry that small sample sizes leave comparative psychology vulnerable to these same problems, the field has only just begun to respond (Beran 2018, Farrar, Boekle, & Clayton 2020). Stevens (2017) notes that comparative psychology makes frequent use of within-subjects methods[6] that might protect the field compared to social psychology.

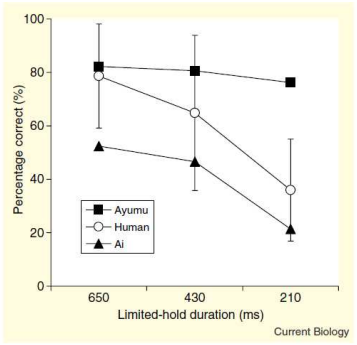


Fig. 2: Data from Inoue & Matsuzawa 2007 (pg. R1005). Two chimpanzees, Ayumu and Ai are compared to a human average.

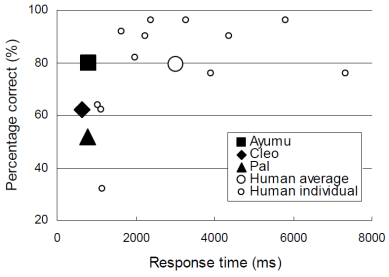


Figure 3: Data from Inoue & Matsuzawa 2007, with 650 *ms* presentation of stimuli before mask (supplemental materials, figure S2). This is the same condition as the leftmost data-points of figure 2, this page.

[5] This is compounded by the fact that Ayumu here is an outlier among even the top performers: only those individuals able to perform the basic task were included, and Ayumu's performance was an outlier among them. There are also concerns that the life-history of laboratory animals makes them unrepresentative.

[6] I note that within-subjects statistical analyses may be more likely replicate even with few individuals, but those methods do not help the problem of generalizing findings to other members of the species.

However, he says, there are several reasons to think that comparative psychology is vulnerable to replication failures. He makes several recommendations for the field to address these concerns. Some of these recommendations have also begun to be implemented. I will focus here on recommendations that inform the current discussion.

One such recommendation is for researchers to pre-register their methods before the test, or for journals to adopt the practice of registered reports, in which a journal accepts or rejects a paper based on methods alone, before experiments are run. This practice has grown in fields like social psychology. The purpose is to prevent fishing-expedition approaches to studies and statistical analyses: These can lead to cherry-picking which studies are reported, and P-hacking by, for instance, simply trying various statistical analyses until one gets a significant result. In 2018, the journal *Animal Behavior and Cognition* began accepting registered reports (Vonk & Kraus 2018), though the editors report that uptake by researchers has been slow (Beran 2020).

Worries about sample size are more complicated. For instance, social psychology has massively increased sample sizes in their studies simply by making greater use of online platforms like Mechanical Turk and Qualtrics. Comparative psychology has no such option. And indeed, for reasons noted above, it seems impossible to completely avoid small sample sizes. Nonetheless, Stevens does make some recommendations that can help. First, different labs can collaborate and combine their subject pool. In fact, the ManyPrimates Project was launched in 2019 to facilitate collaboration across labs spanning the globe, allowing for larger and more diverse samples in studies of primate cognition (Many Primates et al. 2019). Secondly, he suggests that researchers can take advantage of facilities like zoos that may have larger numbers of animals available. Thirdly, researchers can reconsider their choice of species, either by running studies pooling multiple species, or by switching to species that are easily available in the community, such as dogs.

I have little to add on recommendations regarding species choice, but I will take on-board the rest of the recommendations I've mentioned. While the recommendations aimed at increasing sample size are

unlikely to completely address the problem (they simply cannot have an impact like we've seen in social psychology), they certainly help. Registered reports are also valuable; if papers are evaluated based on methods rather than results, it will significantly impact our interpretation of studies will small sample sizes in ways I discuss below (section 6).

Even large-scale changes are not likely to completely address sample size worries in comparative psychology. But even if they do in the future, we should still consider how to interpret existing small sample studies. Either way, interpretive challenges remain. To face these challenges, we can start by look to other research programs that employ very small samples, or even samples of one. To the extent these programs are analogous to comparative psychology, they might provide concrete suggestions.

4. Candidate Analogue One: Cognitive Neuroscience

Lesion studies in cognitive neuroscience present the first candidate analogue. In many of these studies, researchers test a single patient with known brain damage on a battery of tasks aimed at delimiting a certain cognitive capacity.[7] Studies like this generally focus on two kinds of question. The first are questions about the neural underpinnings of a particular cognitive capacity. Here, the goal is locating damage, and correlating it with deficits. The second are the so-called dissociations of capacities that might otherwise be thought to be expressions of a single system. For instance, if a deficit in experiential memory does not also bring with it a deficit in memories for facts, then we have reason to believe that the two are separate capacities subserved be separate systems (episodic and semantic memory), and moreover, the intact capacity does not require the damaged capacity.

The evidential value of lesion studies has long been controversial. As a result, there is a substantial literature aimed at uncovering the methodological assumptions behind the research (e.g. Caramazza 1986, Bub & Bub 1988, McClosky & Caramazza 1988, Glymour 1994, Shallice 2015). The actual damage and

[7] As in the Matsuzawa study, these individuals are also outliers; they are chosen precisely because their performance is abnormal.

deficits observed in individuals vary substantially, and the 'clean' cases of a particular deficit are rare. As a result, it can be difficult to know what aspects of any study can be generalized. Arguably, these concerns, along with improvements in other methods, have driven a reduction in reliance on lesion studies in recent decades. However, if one *is* dealing with lesion studies, the focus on a specific individual is arguably (but controversially) an advantage. The very fact that individual deficits vary so much means that effects would likely wash out in any cohort study, leaving them impossible to interpret (Caramazza 1986).

Even so, there is at least one kind of general claim that these studies do seem to license. These are claims about the *necessity* of one capacity for another, as made in dissociation studies. If Task A can be performed by an individual who cannot perform Task B, then it cannot be the case that the capacity responsible for performance of Task A is necessary for performance on Task B.[8] This inference can be transferred. For instance, the fact that Ayumu was able to do so well on the memory task without using language suggests that language is not required. Necessity claims are strong claims though, especially for a field like psychology, where pretty much everything can vary across individuals. So, the denial of a necessity claim may not always be hugely informative. Nonetheless, even if this is a limited result, it's something.

5. Candidate Analogue Two: Anecdotes in Cognitive Ethology

Researchers in cognitive ethology will also sometimes report anecdotes, or "incident reports" of particular observed behaviors. As with lesion studies, this practice is controversial (Mitchell, Thompson, & Miles 1997). In general, data based on repeated observation is preferred, if possible. Even so, incident reports reports may describe low-frequency behaviors, that would be difficult to observe frequently or to elicit in a laboratory setting. They can also introduce behaviors that researchers had been wholly unaware of. Field anecdotes can also arguably provide some evidence about cognitive processes on their own: field

[8] This basic inference structure is also employed in developmental psychology, though with larger sample sizes (Perner & Lang 1999).

observations don't face any concerns about ecological validity, and anecdotes can often supply richer context about the individual behaving and its context than experiment (Mitchell 1997).[9]

Nonetheless, incident reports do suffer from the limitations described above, with concerns about anthropomorphism and generalizability at the fore. Indeed, the use of anecdotes has been declining in primatology (Ramsay & Teichroeb 2019), suggesting that the downside of anecdotes is winning out in the minds of researchers. Even if these anecdotes do not provide much evidential value, they have heuristic value in generating hypotheses, guiding future observation or experimentation, and identifying behaviors worthy of more systematic study (Silverman 1997¸ Andrews 2020).

6. Anecdotal Experiments

As a start towards coming to grips with the sample size problem in comparative psychology, I argue that we should view studies with extreme small samples sizes as *anecdotal experiments*. Anecdotal experiments have some of the strengths that are usually ascribed to well-designed experiments (they are controlled and meticulously recorded), and some of the weaknesses ascribed to standard anecdotes (they may not be reliably repeatable, and they do not support straightforward generalization to other individuals). They occupy a middle-ground, providing stronger evidence than that provided by a one-off observation, but not as strong as that provided by experiments with larger sample sizes.

To illustrate more specifically, I return to the concerns lodged against anecdotes in section 1. Anecdotal experiments avoid the most significant concerns, while the rest could be lodged against these studies anyway. I'll work through each in turn.

Concern 1: Anecdotes can be cherry-picked to make a predetermined point.

This worry can be avoided by making use of registered reports, such that papers are accepted based on methods, before experiments are done. It remains a worry that existing studies report cherry-picked

[9] Mitchell advocates specifically for anthropomorphic anecdotes as a way to conceptualize behavior. I set the issue of anthropomorphism aside for now, as I see it as less of a concern here (see next section).

experiments, though perhaps not to the degree of full anecdotes: the number of individual behaviors one might observe and dismiss in reporting an anecdote is much less than the number of experiments one might perform and dismiss.

Concern 2: We lack control over and knowledge of background conditions of anecdotes.

This worry does not apply here to any greater degree than it does in psychology generally. A well-designed experiment controls immediate background conditions, such that we can have a reasonable idea of what features of the task the animal is responding to.

Concern 3: Anecdotes are non-repeatable (non-replicable), and so can't be confirmed independently.

Anecdotal experiments have records of methods, which make replication possible. However, replication problems in other areas suggest that comparative psychology should be concerned about replicability (Farrar, Boekle, & Clayton 2020). Perhaps the focus on within-subject tests puts comparative psychology in somewhat better position than it might be otherwise (Stevens 2017), but the extremely small sample sizes suggest that replicability cannot be assumed. This is a worry either way, and framing these as the anecdotal experiments can make it more explicit.

Concern 4: Anecdotes are narrative in structure, rather than providing analyzable data.

Anecdotal experiments do rely on data, so seem to pass this test. Nonetheless, we should be careful in what we take that data to show. If, as just suggested, we should question the replicability of these studies, statistics can mislead. A careful reevaluation of statistical measures can help here (as in social psychology). However, absent that, statistics can present a false sense of generalizability. For instance, we can statistically show that Ayumu himself reliably outperforms the human sample average in this study. What that means about chimpanzees more generally is a different question.

Concern 5: They don't support generalization.

As with concern 3, this is just to recognize limits already present. It is common to restate an experimental finding by simply plugging generics into a literal description of the study. For instance: "Ayumu outperformed the average performance of twelve humans in our study" becomes "chimpanzees outperform humans." This move is clearly too quick. If we have good reason to believe that a study includes a representative sample of a larger population, we generalize to that population. These generalizations should become more tentative as confidence in the representativeness of the sample increases. With very few animals, we can't generalize this way. This is compounded by the fact that the animals performing in these experiments, like Ayumu, are often outliers.

Treating experiments with extremely small sample sizes as anecdotal experiments marks their limitations, and helps guide their proper use. There are many important unanswered questions here. We would want to know how to determine which experiments are anecdotal and which are not; where is the cut-off? Moreover, in light of the interpretive limitations of anecdotal experiments, I have said little about what, concretely, we can learn from them. I will offer some brief comments on that topic here.

The fact that one member of a species is able to perform a task to a certain criterion shows that it is *possible* for some members of that species to do so. However, this doesn't guarantee any particular cognitive mechanism. Though, we can follow work in cognitive neuroscience and conclude that successful performance shows that some capacity believed to be absent (say, language) is not *necessary* for performance on the task. They may also provide some evidence for one hypothesized mechanism over another if that level of performance is impossible or highly implausible according to the devalued hypothesis. Absent such strong claims, one competing hypothesis may still predict better performance on a task (this is not the aim of the Inoue & Matsuzawa study). If so, a convincing finding of strong performance might provide a small (minute, even) amount of evidence for that hypothesis. Additionally, following cognitive ethology, the fact that at least one individual succeeds in a task might motivate new hypotheses about the cognitive capacities involved, or identify new areas worthy of further study. These are useful conclusions, but they are not often deeply helpful in evaluating models of the actual cognitive

processes involved. Psychological models rarely make claims of possibility or impossibility, and one cannot conclude a capacity is not necessary for the task unless one is confident the animal does not possess that capacity (most of the interesting options are still up in the air).

Even with these limitations in scope and strength, any generalization from an extremely small sample to a species at large must be significantly hedged: these individuals might just be doing something completely different than other members of the species.[10] But even so limited, there is still value to that evidence. Often when it comes to nonhuman minds, strong evidence is very hard to come by, so any amount of evidence is worth considering.

7. Implications and Conclusion

The basic point of framing extreme small sample studies as anecdotal experiments is to reduce their weight in general claims about the nature of nonhuman cognitive capacities. Indeed, I argue that the field ought to reduce the evidential weight of individual experiments in general, to help move away from a pernicious 'critical experiment' framing that still too often pervades. The actual evidential value of individual experiments must be assessed on a case by case basis, depending on the kind of model being evaluated, and the nature of the anecdotal experiment. This is tough work of course, but it always has been.

There may be other general impacts on the field. This framing could benefit the field by encouraging more exploratory research and reporting of more varied behaviors. In effect, experimental comparative psychology might look a bit more like field ethology. For example, Stanton et al. (2017) presented raccoons with the Aesop's fable task, in which they can gain access to a treat floating on water by dropping stones in to raise the water level. They report that one of the raccoons managed to get the treat, not by dropping stones, but by ripping the entire apparatus off the floor and dumping it out. A field that

[10] I have ignored worries about ecological validity and differences between captive and wild animals, but they would have to be considered in addressing this possibility.

relies on registered reports, and recognizes the limitations of data from such small sample sizes would likely include substantially more reports of behavior like this. There is value to that, as these behaviors, intended by the experimenter or not, do provide insight into the animals.

Most importantly, though, this framing encourages more honest reporting of the significance of studies. Extreme low sample size studies are limited in evidential value. Reporting them as anecdotal experiments presents them as such.

References

Andrews, K. (2020). *How to Study Animal Minds*. Cambridge: Cambridge University Press.

Beran, M. (2018). Replication and Pre-Registration in Comparative Psychology. *International Journal of Comparative Psychology*, *31*.

Beran, M. (2020). Editorial: The Value and Status of Replications in Animal Behavior and Cognition Research. *Animal Behavior and Cognition* 7(1): i-iii.

Bub, J. and Bub, D. (1988): On the Methodology of Single-case Studies in Cognitive Neuropsychology, *Cognitive Neuropsychology*, 5, 563-582.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, 5, 41–66.

Dacey, M. (2017). Anthropomorphism as cognitive bias. *Philosophy of Science*, *84*(5), 1152-1164.

Farrar, B. G., Boeckle, M., & Clayton, N. S. (2020). Replications in comparative cognition: What should we expect and how can we improve? *Animal Behavior and Cognition*, 7(1), 1-22. doi: https://doi.org/10.26451/abc.07.01.02.2020

Inoue, S., & Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. *Current Biology*, *17*(23), R1004-R1005.

Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., ... & Flessert, M. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PloS one*, *14*(10).

McCloskey, M., & Caramazza, A. (1988). Theory and methodology in cognitive neuropsychology: A response to our critics. Cognitive Neuropsychology, 5, 583–623.

Mitchell, R. W. (1997) Anthropomorphic Anecdotalism as Method, in Mitchell, R. W., Thompson, N. S. & Miles, H. L. (Eds.). Anthropomorphism, Anecdotes, and Animals. Albany: State University of New York Press pp 151-169.

Mitchell, R. W., Thompson, N. S. & Miles, H. L. (1997). Anthropomorphism, Anecdotes, and Animals. Albany: State University of New York Press.

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in cognitive sciences*, *3*(9), 337-344.

Ramsay, M. S., & Teichroeb, J. A. (2019). Anecdotes in Primatology: Temporal Trends, Anthropocentrism, and Hierarchies of Knowledge. *American Anthropologist*, *121*(3), 680-693.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, *14*(11), e12633.

Shallice, T. (2015). Cognitive neuropsychology and its vicissitudes: The fate of Caramazza's axioms. *Cognitive neuropsychology*, *32*(7-8), 385-411.

Shettleworth, S. (2012). Fundamentals of Comparative Cognition. Oxford: Oxford University Press.

Silverman, P. S. (1997). A Pragmatic Approach to the Inference of Animal Minds, in Mitchell, R. W., Thompson, N. S. & Miles, H. L. (Eds.). Anthropomorphism, Anecdotes, and Animals. Albany: State University of New York Press pp 170-185.

Stanton, L., Davis, E., Johnson, S., Gilbert, A., & Benson-Amram, S. (2017). Adaptation of the Aesop's Fable paradigm for use with raccoons (Procyon lotor): considerations for future application in non-avian and non-primate species. *Animal cognition*, *20*(6), 1147-1152.

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in psychology*, *8*, 862.

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. New York: The MacMillan Company

Vonk, J., & Krause, M. A. (2018). Editorial: Announcing preregistered reports. *Animal Behavior and Cognition*, *5*(2).

STRUCTURAL HUMILITY

We can see our world (and possible worlds generally) as naturally dividing up into structure and contents. The contents of the world further divide into the properties and individuals which are instantiated at and exist in the considered world, respectively. On the other hand, the structure of the world provides the way that the contents are organized.

Call a thesis a 'Humility Thesis' if it amounts to claiming that there is some important part of the world that we are irremediably ignorant of. Humility Theses are claims of some systematic epistemic limitations we have. For example, David Lewis, in his "Ramseyan Humility" (2009), argues that we are irremediably ignorant of the identities of many properties of things.[1] We only come to know them as role-occupants (of dispositions or other roles). But given a contingent connection between roles and occupants, different properties can occupy the same role at different worlds. Thus, knowledge that the role is occupied is insufficient for identifying the property occupying that role. An analogous Humility Thesis arises in the case of individuals. Assume that we can know the qualitative character the individuals in our world. If individuals are only contingently connected with their qualitative properties (in the way role occupants were suggested to be connected with their roles), then different individuals could occupy the same qualitative characters in different worlds. If this is so, then knowledge that a particular qualitative character is had is likewise insufficient to know the identity of the individual which has that character.

The routes just sketched for these two Humility Theses bear a significant similarity. Both aforementioned Humility Theses involve claims about our epistemic limitations regarding our knowledge of the identities of *contents* of the world. The question I want to ask in this paper is whether or not there is reason to think that we may be irremediably ignorant of the *structure* of the world. Spatiotemporal structures provide common examples of world structures. As such I will limit the following discussion to whether or not we should accept a Humility Thesis about the world's spatiotemporal structure.[2] In particular, I argue that we remain irremediably ignorant of whether we are in a world with distinct regions which are topologically indistinguishable from one another.

[1] I discuss Lewis' arguments in §1.

[2] In what follows I ignore current discussion about whether or not our world isn't fundamentally spatiotemporal, though I believe that my discussion will generalize to other types of world structures.

I begin by briefly reviewing Lewis's argument for Humility about the intrinsic properties of things ('Ramseyan Humility' henceforth). I then discuss whether we should endorse a corresponding Humility Thesis about the worlds' spatiotemporal structure ('Structural Humility' henceforth). I argue that the standard metaphysics of spacetime fall prey to Structural Humility. This is significant because avoiding concerns of Humility is touted as a reason for adopting a particular metaphysic of spacetime. I conclude with a brief discussion of the implications of Structural Humility for this view.

1. RAMSEYAN HUMILITY

Lewis's argument for Humility regarding our knowledge of properties begins with two arguments for Humility about *fundamental* properties.[3] Fundamental properties can come in a variety of categories as well. They can be all-or-nothing properties of various adicities, or come in varying degrees such as scalar and vector magnitudes, and so on.

Advances in scientific theorizing and the discovery of fundamental properties stand in a mutual relationship. So much so that a true and complete final theory, $\mathscr{T}$, will provide us with a complete inventory of the fundamental properties at work in nature. The final theory, $\mathscr{T}$, however, will leave out properties which are instantiated but play no role in nature ('idlers'), and those fundamental properties which aren't instantiated in our world ('aliens').[4]

The argument for Ramseyan humility can be seen as proceeding in two steps. First, the argument shows that any evidence for our fundamental theory $\mathscr{T}$, is just evidence for what is called the *Ramsey sentence* of $\mathscr{T}$. Second, it is argued that the Ramsey sentence of $\mathscr{T}$ admits of multiple realizations. Since all evidence for $\mathscr{T}$ is only evidence for the Ramsey sentence of $\mathscr{T}$ and the Ramsey sentence of $\mathscr{T}$ isn't uniquely realizable, we have no more evidence for $\mathscr{T}$ than any other possible realizer of the Ramsey sentence of $\mathscr{T}$. Allow me to unpack.

Recall $\mathscr{T}$ is our final and complete theory at the limit of empirical enquiry. The language of $\mathscr{T}$ contains $\mathscr{T}$*- terms* which are the theoretical terms implicitly defined by $\mathscr{T}$. Then there is the rest of our language which Lewis calls $\mathscr{O}$*-language* for 'old language'. $\mathscr{O}$ -

[3] Lewis (2009, pp. 204-5) tells us that the fundamental properties are those that ground objective similarity and difference, they provide a minimal base for the rest of the world's qualitative features. For more in-depth treatments of fundamental properties see Lewis (1983) and Lewis (1986, pp. 59-63).
[4] Lewis (2009, p. 205).

language is what is available to us without the term introducing theory $\mathscr{T}$. The $\mathscr{O}$-language is rich enough to describe all possible observations.[5]

Recall that all fundamental properties except aliens and idlers will be listed in $\mathscr{T}$'s inventory. Importantly, all of the fundamental properties mentioned in $\mathscr{T}$ are named by the $\mathscr{T}$- terms.[6] Now the theory $\mathscr{T}$ consists in all of the logical consequences of a sentence called the *postulate* of $\mathscr{T}$. We can write the postulate as $\mathscr{T}(t_1, \ldots, t_n)$ where $t_1, \ldots, t_n$ are the theoretical terms introduced by $\mathscr{T}$ and all of the rest of the language in the postulate is $\mathscr{O}$-language. When we replace all of the $\mathscr{T}$- terms with variables, we get $\mathscr{T}(x_1, \ldots, x_n)$. An *n*-tuple that satisfies $\mathscr{T}$ with respect to the actual world is called an *actual realization* of $\mathscr{T}$ whereas one that can satisfy $\mathscr{T}$ with respect to some possible world is a *possible realization* of $\mathscr{T}$. We then get the *Ramsey sentence* of $\mathscr{T}$ when we prefix $\mathscr{T}(x_1, \ldots, x_n)$ with existential quantifiers: $\exists x_1, \ldots, \exists x_n, (x_1, \ldots, x_n)$.[7] Significantly, the Ramsey sentence of $\mathscr{T}$ implies exactly those $\mathscr{O}$- language sentences which are implied by the postulate of $\mathscr{T}$.[8] Because the $\mathscr{O}$- language is rich enough to describe all possible experiences, the predictive success of $\mathscr{T}$ will be the same as the Ramsey sentence of $\mathscr{T}$. This means that if there are multiple possible realizations of the Ramsey sentence of $\mathscr{T}$, no possible observation can tell us which one is the actual realization. This is because, no matter which one is the actual realization, the Ramsey sentence will be true and our observational evidence only gives us evidence for the truth of the Ramsey sentence.[9]

What is left to be shown is that there are in fact multiple realizations of the Ramsey sentence of $\mathscr{T}$. Lewis offers two arguments for this conclusion: the *permutation argument* and the *replacement argument*. Both rely on Lewis's acceptant of a principle of recombination. Namely, that we can take apart distinct elements of a possibility and rearrange them, we can remove some of the distinct elements, we can reduplicate some of them, and we can replace elements of some possibility with elements of others and get a new possibility.[10] It is

[5] *Ibid.* pp. 205-6.
[6] *Ibid.* p. 206.
[7] *Ibid.* p. 207.
[8] *Ibid.* p. 207, n. 6.
[9] *Ibid.* p. 207.
[10] *Ibid.* pp. 207-8. For an in-depth discussion into formulating a principle of recombination and other principles of plenitude see Bricker (MS b).

important to note that distinct elements cannot be recombined in any way possible, but that they have to be recombined in a category-preserving way.

The permutation argument starts with the assumption that we have the actual realization of $\mathscr{T}$. Then we find the members of the *n*-tuple that satisfies $\mathscr{T}$ that are fundamental and belong to multi-membered categories. Then we permute these within their categories to get a new *n*-tuple that satisfies $\mathscr{T}$. The principle of recombination is what allows us to permute these properties to get a possibility. *Quidditism*, the view that two worlds can differ merely by permutation of fundamental properties, gets us that the resulting possibility is distinct from the original possibility. Note that the argument from permutation only gets us humility insofar as there are actual fundamental properties of multi-membered categories that can be swapped. If there are only a small number of categories of fundamental properties in $\mathscr{T}$ that are multi-membered, this does not guarantee a sweeping Humility Thesis.[11] The replacement argument is designed to provide a more sweeping conclusion.

The replacement argument gets us Humility through replacing the fundamental properties in $\mathscr{T}$ with fundamental alien and idling properties of the same category. If there are alien or idling properties that fall into the same categories as the fundamental properties mentioned in $\mathscr{T}$, then recombination entails that there are distinct possibilities where some or all of the fundamental properties in $\mathscr{T}$ have been replaced with aliens or idlers of the same category. Lewis offers a few reasons to think that there will be enough alien properties to replace at least a large majority of the fundamental properties in the actual realization of $\mathscr{T}$. The reason I find the most powerful begins by noting that it is a contingent matter what fundamental properties are instantiated. And once we've appreciated this fact we should think that there is a world where more fundamental properties are instantiated than are instantiated at this world. And there is a further world with more properties instantiated at it than the second one and so on. It's implausible to think that amongst these worlds with more fundamental properties than ours that there won't be alien properties that are members of most, if not all, of the categories in the fundamental properties mentioned in $\mathscr{T}$. Thus, we have good reason to think that there are sufficiently enough alien properties

[11] Lewis (2009, 208-12).

for the replacement argument to go through. This argument gives us an argument for a much more sweeping Humility Thesis than the permutation argument.[12]

2. HUMILITY ABOUT SPATIOTEMPORAL STRUCTURE

We've seen how Lewis argues for a Humility Thesis about our knowledge of the properties our world instantiates in his arguments for Ramseyan Humility. To get to Structural Humility we need to proceed differently. One important reason for thinking this has to do with the inapplicability of recombination to structure. Lewis's arguments for Ramseyan Humility made use of recombination to swap properties around from within a world or swap properties from a different world into the structure of the old one in order to get new possibilities. We can't swap around parts of structures in the same way. Trying to use recombination to fill out the possible world structures runs into serious problems. Further, the recombination principle that Lewis uses presupposes that there is a structure to recombine the elements into. Instead, we need a different principle of plenitude for structures. I believe if we accept a plausible principle of plenitude for world structures *and* we accept some plausible views about the nature of the worlds' geometric structure, then we remain irremediably ignorant of important aspects of the worlds' geometric structure. Namely, whether the world we live in contains distinct topologically indistinguishable points and how the distinct indistinguishable points are distributed.[13]

2.1. Metric and Merely Pseudo-Metric Spaces

First, let's take a look at two different classes of geometric structures. *Metric spaces* are spaces whose topology is solely determind by a distance function that meets the following definition:

D1 $d(x,y) = 0 \Leftrightarrow x = y$	(Identity of Indiscernables)
D2 $d(x,y) = d(y,x)$	(Symmetry)
D3 $d(x,y) + d(y,z) \geq d(x,z)$	(Triangle Inequality)

3D-Euclidean spaces count as an examples of a metric space. The metric spaces are part of the larger class of *pseudo-metric spaces*. That is all metric spaces are pseudo-metric spaces but not all pseudo-metric spaces are metric spaces. The class of pseudo-metric spaces is the class of spaces whose topology is defined by a distance function that replaces D1 with:

[12] *Ibid.* 212-4.

[13] Any distinct points, p and p^*, are *topologically indistinguishable* just in case for any open set, S, p belongs to S just in case p^* belongs to S.

D1* $x = y \Rightarrow d(x,y) = 0$ (Indiscernibility of Identicals)

In other words pseudo-metric spaces include geometric structures which have distinct points at zero-distance from one another. Call the pseudo-metric spaces which have distinct points at zero-distance from one another *merely pseudo-metric spaces*. The metric spaces and the merely pseudo-metric spaces are mutually exclusive and exhaust the class of pseudo-metric spaces. Metric spaces and merely pseudo-metric spaces only differ over whether they have *topologically indistinguishable* points or not. In metric spaces the open sets that fix the topology also uniquely determine the points in that space, in merely pseudo-metric spaces this is not the case. Moreover, in merely pseudo-metric spaces there will also be distinct topologically indistinguishable regions besides the point-sized ones. For any two distinct topologically indistinguishable regions, R and R^*, there are some distinct topologically indistinguishable points, p and p^*, such that p is in both R and R^* yet p^* is in R but not R^* (or *vice versa*).

2.2. The Possibility of Merely Pseudo-Metric Spaces

We are accustomed to thinking in terms of spaces that are metric spaces. In fact, I'd imagine most think it is constitutive of being a point that it is uniquely identified by its place in the worlds' geometric structure. The possibility of merely pseudo-metric spaces flouts this intuition. So there needs to be good reason to think that merely pseudo-metric spatial structures are possible. The best way to go about this requires providing a principled way to determine what structures are possible and which ones aren't. In "Plenitude of Possible Structures" (MSa) Bricker provides what I take to be the best method for determining the possibility of a class of world structures.

The method can be summed up as follows: First, we need to determine what structures have played an explanatory role in our theorizing about the world.[14] Here, playing an explanatory role isn't understood in sociological, but objective terms – the structures must have genuine explanatory power.[15] Determining these structures provides the base of logically possible structures from which we can generalize to other possible structures. Next, we need to determine which classes of structures are natural classes. The members of

[14] In particular Bricker tells us "[w]e have warranted belief that a structure is logically possible if that structure plays, or has played, an explanatory role in our theorizing about the actual world." (MSa, p. 5). This just gives us a base set of structures from which we will determine the who class or classes of possible structures from.
[15] Bricker (MSa, p. 6).

natural classes of structures objectively resemble each other in ways that members of classes that aren't natural don't. We determine the natural classes of structures by seeing whether or not each of them serve as a principle object of study in some major area of study in mathematics – the ones that do are the natural classes.[16] This gives us candidate natural classes of structures to generalize to as logically possible ones. Finally, not just any generalization from the base classes of structures to a natural class will count as a good generalization. Only those natural classes that are *natural generalizations* of the structures in our base count as logically possible structures.[17] Here, again, we defer to mathematicians to see what classes of structures are natural ganeralizations of others. This method gives us the following principle of plenitude:

> PRINCIPLE OF PLENITUDE OF STRUCTURES
> Suppose $\mathcal{S}$ is a class of logically possible structures. Any structure belonging to any natural generalization of $\mathcal{S}$ is logically possible.[18]

The argument for the possibility of merely pseudo-metric structures is straightforward. First, the class of Euclidean spaces, $\mathscr{E}$, is a prime example of a class of structures that have played a role in our theorizing about the actual world. So $\mathscr{E}$ is a class of logically possible structures. The class of metric spaces, $\mathscr{M}$, is a natural generalization of $\mathscr{E}$, so this means any structure in $\mathscr{M}$ is logically possible. This is the same as saying that $\mathscr{M}$ is a class of logically possible structures. Finally, the class of pseudo-metric spaces, $\mathscr{P}$, is a natural generalization of $\mathscr{M}$. Because $\mathscr{M}$ is a class of logically possible structures, and $\mathscr{P}$ is a natural generalization of $\mathscr{M}$, any structure in $\mathscr{P}$ is logically possible. All of the structures of pseudo-metric spaces are in $\mathscr{P}$. This includes all of the merely pseudo-metric spaces. So, merely pseudo-metric spaces are logically possible. Moreover, *any* merely pseudo-metric spatial structure is a logically possible one.

2.3. Undetectible Differences

Recall that Lewis's argument for Ramseyan Humility is intended to show that although we can come to know the properties of things as role-occupants this is insufficient to identify

[16] Bricker (MSa, pp. 9-10). Though, they aren't natural *because* they are the principle objects of study some major area in mathematics. Instead, they are the principle objects of study in some major area in mathematics *because* they are natural.
[17] Bricker (MSa, p. 19).
[18] Bricker (MSa, p. 19).

the role-occupier. The points, specifically, and regions, generally, in a geometric structure can likewise be thought of as role-occupants of that particular geometric structure. I would like to suggest that we can think of the difference between merely pseudo-metric and metric spaces in a similar way. The rough thought is that the distinct but topologically indistinguishable points in merely pseudo-metric spaces play the same role as the unique topologically distinguishable points in metric spaces. To put the idea slightly differently, we can't tell how complex the occupants of the point-roles in the worlds' geometric structure are. This isn't quite right, but provides us with a useful, albeit imperfect, way of drawing out the similarity between Ramseyan Humility and Structural Humility.

An important feature of merely pseudo-metric spaces is that there is a way to "convert" them into metric spaces. Recall that the only difference between a particular metric spatial structure and its equivalent merely pseudo-metric structures is that they disagree on whether or not there are distinct topologically indistinguishable points. Different merely pseudo-metric structures that are otherwise structurally the same as a given metric space will only differ on how many distinct indistinguishable points there are and the distribution of the distinct topologically distinguishable points. This could be as minimal of a difference from the corresponding metric space as there being exactly two points in a merely pseudo-metric structure that are topologically indistinguishable to all of the points being topologically indistinguishable from some other distinct points. Some pseudo-metric spaces may uniformly increase the topologically indistinguishable points, so that for each distinguishable point in the metric space, there are 2, or 3, or 4, … indistinguishable points in the merely pseudo-metric space. Or, the increase could be non-uniform. Nevertheless, each of these merely pseudo-metric spaces can be converted into metric spaces by treating the pluralities, or fusions, or sets of distinct topologically indistinguishable points in a merely pseudo-metric space as single points in a metric space.

To see this, let X be a merely pseudo-metric space. Let $x \sim y$ just in case $d(x, y) = 0$ (*i.e.* just in case x and y are topologically indistinguishable in X). So any points stand in the equivalence relation '$\sim$' if they are zero distance from each other according to the distance function, d, defined on X. We can then define a new space X^* where $X^* = X/\sim$. In the new space, X^*, each of the points are equivalence classes of points in X, represented as $[x]$, $[y]$. We define a distance function d^*: $X/\sim \times X/\sim \rightarrow \mathbb{R}_+$ such that $d^*([x], [y]) = d(x, y)$. We can see that d^* is a metric and X^* is a metric space. For, we already know that d^* will satisfy D1*, D2 and D3 of the definition of a metric above, and, further, because $x \sim y$ if and only

if $d(x, y) = 0$, then $d^*([x], [y]) = 0$ if and only if $[x] = [y]$. So D1 will be satisfied. The space X^* is called the *metric identification* of X.[19]

Through metric identification, it seems like *any* theory that is cast in terms of a metric structure could be cast in terms of a merely pseudo-metric structure. Where the metric theory has simple, singular, and distinguishable points filling the roles of the point-sized regions, the merely-pseudo-metric theory will have pluralities of distinct indistinguishable points or their fusions filling these roles. Further, no matter which way the world turned out it seems we would be none the wiser. The pluralities of distinct indistinguishable points in the pseudo-metric version of the theory will do the same work in the theory's predictions as will the single distinguishable points in the metric version of the theory. Same predictive work, same amount of confirmation. If this is right, then there is an important part of the worlds' geometric structure we will remain forever ignorant of.

Now, I'd imagine that one might want to object that we would *have no reason* to posit the extra indistinguishable points that the pseudo-metric version of the theory does. This is because simplicity dictates that we should accept the simpler of the two versions of the theory. Because the pseudo-metric version of the theory makes unnecessary posits, then we should prefer the metric version of the theory. I do not find this objection compelling. We are interested in what we can know. If the sense of 'prefer' here has to do with knowledge, then the objector has to tell us how we could know that the world is simpler in this way. But, this is *just what I've argued we couldn't do.* Perhaps they might say that we could, in principle, build some detection device that could detect whether or not indistinguishable points or regions were present. Assume that one could build such a device. This device would have to operate based off of some sort of *causal connection* with the distinct indistinguishable regions that allowed it to detect when multiple regions take the same position in spacetime. Even if this were possible, this would still leave undetermined important facts about the worlds geometric structure. Any theory, T, by which our detection device would work, would have to spell out what the causal conditions were whereby it would be able to detect the presence of multiple indistinguishable points. Note that theory T will only distinguish topologically distinguishable points by the causal role that they play. So we only come to know and identify the points by the causal role they play. Now, imagine a different theory, T^*, which is identical to T *except* that whenever the causal roles are filled that allows us to detect the presence of distinct topologically

[19] For a more thoroughly spelled out version of this proof see Simon (2015, pp. 3-4).

indistinguishable points, according to *T*, in *T** that role is filled by pairs of topologically indistinguishable points. Our detection device would operate in much the same way, and would be able to detect some instances of distinct topologically indistinguishable regions, but it would be none the wiser as to whether it was in a *T* world or a *T** world. The thrust of the idea here is that if we have a theory that makes some claim about the points and how they are distinguished, we can replace it with a theory where pluralities or complexes of points of whatever number are playing those exact same roles. Because of this we will forever remain unable to know important features of our worlds' geometric structure.

It is important to notice that this argument takes seriously the idea that the spatiotemporal structure of the world includes something like points. How seriously must we take the existence of points to get Structural Humility off of the ground? Not very, I think. There are three major contenders in the debate over the nature of spacetime: substantivalism, ontic-structural realism ('structuralism' henceforth), and relationalism. None escape Structural Humility. Let me briefly explain why. Substantivalists of all stripes take the worlds' spacetime to be fundamental, independent thing. This means the substantivalist takes regions and the spacetime structure as fundamental. Substantivalists will agree that spacetime is made up of points connected in a structure of spatiotemporal relations. Since points are genuine objects according to the spacetime substantivalist, the world structures that are strictly pseudo-metric will be understood in terms of real, physical, distinct topologically indistinguishable points and the threat of Structural Humility will loom. Structuralists, on the other hand, don't take points very seriously at all. For them, the spatiotemporal structure is fundamental, and the points are, at best, placeholders in the structure lacking intrinsic natures, and, at worst, just places or intersections in the series of relations that constitute the worlds' spatiotemporal structure.[20] However, structuralists still have to worry about Structural Humility. Roughly, the structuralist maintains the relational structure posited by the substantivalist but loses the points.[21] So a world with a pseudo-metric structure, for the structuralist, will have distinct indistinguishable places within its structure. How many of these there are, or how they're distributed will remain forever unknown to us. Finally, the relationalist takes the worlds' spatiotemporal structure to be dependent upon the material objects and the fundamental spatiotemporal relations they

[20] See Esfeld and Lam (2007) for an overview about ontic structural realism and a defense of a moderate structuralism about spaceime.

[21] For example, see Esfeld and Lam (2008, pp.42-3).

stand in. For the relationist the problem arises when we have co-located material objects that are constantly adjoined throughout their existence.

3. CONCLUDING THOUGHTS: STRUCTURALISM AND HUMILITY

So far we've reviewed how Lewis argued for our irremediable ignorance of the identities of many of the properties in the world and I've argued that a similar Humility Thesis about the geometric structure of the world can be seen to follow from some important ontologies of spacetime. The worry was that we are irremediably ignorant of the existence and distribution of indistinguishable regions. This kind of Humility afflicted both substantivalists and structuralists about spacetime but not relationalists. Before closing the paper I would like to briefly note how Structural Humility relates to a kind of strategy that has been used to motivate structuralism.

Structuralism about spacetime is of a piece with a broader ontic structural realist project which seeks to downplay the importance of objects and inflate the importance of structure. Structure is generally treated as being fundamental and objects are taken to be eliminated, reduced to, grounded in, or dependent upon fundamental structure.[22] One important motivation for structuralism is the following kind of consideration:

> **EPISTEMELOGICAL-ONTOLOGICAL COHERENCE (EOC)**
> Our metaphysics should be coherent with our epistemology. Metaphysics that posit entities that lead to unknowable gaps between our metaphysics and epistemology should be done away with. We shouldn't deny ourselves in principle epistemic access portions of (physical) reality. Only structuralism avoids a metaphysics which entails epistemic gaps.[23]

Other motivations for structuralisms in various areas of ontology exploit similar considerations.[24] Motivations, like EOC, can just be seen as denials of a particular Humility Thesis. In the case of EOC the denial of Humility is broad and global. So, if this kind of motivation for structuralism holds water, then structuralism better be able to avoid Humility Theses of any variety. However, if what I've said above is right, then structuralism

[22] See Frigg and Votsis (2011) for a wonderful overview of the varieties of ontic structural realism. Ladyman (1998) and French (1998), depending on how they are read, can be seen as advocating either an eliminative or reductionist approach. MacKenzie (2018) provides a grounding based understanding of ontic structural realism. And Esfeld and Lam (2008) and Mackenzie (2013) offer versions of structuralism where the relation between objects and structures should be understood in terms of dependence that isn't grounding.

[23] This formulation roughly follows Esfeld and Lam (2008, p. 30). See also Esfeld (2004, pp. 614-6),

[24] See for example, Jantzen's (2011, pp. 435-9) discussion of how standard or naïve realism falls prey to worries about making our physical theories incomplete while structuralism avoids this problem.

cannot avoid Humility across the board – it runs into Structural Humility. As such, considerations about of Structural Humility undercut one important motivation for the ontic structural realist project.

REFERENCES

Bricker (MSa) Plenitude of Possible Structures

Bricker (MSb) Principles of Plenitude

Esfeld, M. (2004). Quantum entanglement and a metaphysics of relations. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *35*(4), 601-617.

Esfeld, M., & Lam, V. (2008). Moderate structural realism about space-time. *Synthese*, 160(1), 27-46.

French, S. (1998). On the whithering away of physical objects. In Castellani, E. (1998). *Interpreting bodies: classical and quantum objects in modern physics*. Princeton University Press.

Frigg, R., & Votsis, I. (2011). Everything you always wanted to know about structural realism but were afraid to ask. *European journal for philosophy of science*, *1*(2), 227-276.

Jantzen, B. C. (2011). No two entities without identity. *Synthese*, *181*(3), 433-450.

Ladyman, James (1998). What is structural realism? *Studies in History and Philosophy of Science Part A* 29 (3):409-424.

Lewis, David. "New work for a theory of universals." *Australasian Journal of Philosophy* 61, no. 4 (1983): 343-377.

Lewis, David K. *On the plurality of worlds*. Vol. 322. Oxford: Blackwell, 1986.

Lewis, D. (2009). *Ramseyan humility*. In Braddon-Mitchell, D., & Nola, R. (Eds.). (2009). *Conceptual analysis and philosophical naturalism*. MIT Press.

McKenzie, K. (2013). Priority and particle physics: Ontic structural realism as a fundamentality thesis. *The British Journal for the Philosophy of Science*, *65*(2), 353-380.

McKenzie, K. (2018). Structuralism in the Idiom of Determination. *The British Journal for the Philosophy of Science*, *0*, 1-26.

Simon, Barry. Real Analysis: A Comprehensive Course in Analysis, Part 1. American Mathematical Soc., November 2015.

Climate Models and the Irrelevance of Chaos

Corey Dethier

(Draft – please do not quote without permission)

Abstract

Philosophy of climate science has witnessed substantial recent debate over the existence of a dynamical or "structural" analogue of chaos, which is alleged to spell trouble for certain uses of climate models. In this paper, I argue that the debate over the analogy can and should be separated from its alleged epistemic implications: chaos-like behavior is neither necessary nor sufficient for small dynamical misrepresentations to generate erroneous results. I identify the relevant kind of kind of sensitivity with a kind of safety failure and argue that the resulting set of issues has different stakes than the extant debate would indicate.

1 Introduction

To make predictions about the future of a system, we need to know two things: the initial conditions, or, present state of the system; and the dynamics of the system, or, how it evolves with time. Chaotic systems present particular difficulties because small differences in initial conditions amplify into large differences in the end state of the system. Is there an analogous dynamical property of systems? Intuitively, it seems like there might be: small differences in the dynamics amplify into large differences in the end state of the system.

In a series of papers, a group of philosophers and scientists have argued this analogous dynamical property and that it spells epistemic trouble for certain hypotheses in climate science. Specifically, they argue that because most climate models heavily idealize the dynamics of the climate, the possibility that such models exhibit a dynamical phenomenon analogous to chaos should cause us to have low confidence in the accuracy of quantitative predictions that rest on them. An opposed group of critics have argued that the analogy breaks down and that the epistemic conclusions don't follow; the possibility of small dynamical errors doesn't undermine the general warrant for quantitative climate predictions.

While the mathematical question concerning the alleged analogy with chaos is interesting on its own terms, the focus on it is misleading from a purely epistemic perspective: a tight analogy to chaos is neither necessary nor sufficient for the kinds of epistemic error that motivate the debate. Chaotic behavior involves growth in physical distance with time; such growth is relevant to the accuracy of a given prediction only when (a) the starting distances are small relative to the desired level of precision, (b) the later distances large from the same perspective, and (c) the time frame covered by the prediction the same as that on which the system is chaotic. The type of *epistemic* sensitivity relevant to error is better captured by the failure of a kind of safety condition. And while it is true that there is good reason to worry about safety failures in climate science, the arguments in question are better seen as explaining known safety failures than as providing evidence for the existence of unknown ones.

A more detailed outline of my arguments is as follows. In sections 2 and 3, I briefly characterize the debate over dynamical analogues of chaos and argue that it has misfired insofar as it presupposes a connection between chaos-like behavior and (the probability of) error. There's no real connection here because chaos involves a type of interest-independent sensitivity, whereas the probability of error is inherently dependent on our interests. In section 4, I provide an alternative notion of sensitivity that is appropriately interest-dependent. This notion is best expressed in terms of the failure of a kind of safety principle—essentially, the safety principle fails when a hypothesis is only justified given an assumption that's uncertain or risky. Finally, in section 5, I offer a reinterpretation of the original arguments: what they motivate is low confidence in our ability to substantially increase the precision of model reports.

2 The debate over dynamical analogues of chaos

The debate over dynamical analogues of chaos has largely focused on a particular minimal condition on chaotic behavior, what's known as "sensitive dependence on initial conditions" (SDIC). Roughly speaking, a system exhibits SDIC if "even arbitrarily close initial conditions will follow very different trajectories" through the state space that characterizes the system (Frigg, Bradley, et al. 2014, 34). Trivially, SDIC implies that a model that slightly misrepresents the initial conditions of the system will misrepresent (some) later states of the system to a much larger degree. In a series of recent

papers, a group of philosophers and scientists associated with the London School of Economics—and thus termed the "LSE group" by their critics—have argued that "structural model error" (SME) presents epistemological problems similar to those presented by SDIC (Frigg, Bradley, et al. 2014; Frigg, L. A. Smith, and Stainforth 2013, 2015).[1]

To be precise, it's important to recognize that SME is *not* supposed to be directly analogous to SDIC. On the contrary, SME occurs whenever a model misrepresents the dynamics of the target system (Frigg, Bradley, et al. 2014, 32). The analogy, according to the LSE group, is between the behavior of non-linear systems relative to SME and to the behavior of SDIC-exhibiting systems relative to misrepresentation of initial conditions. That is: the existence of small amounts of SME will lead a model to dramatically misrepresent some later states of a non-linear system in much the same way that a small misrepresentation of initial conditions will lead a model to dramatically misrepresent some later states of a chaotic system. In effect, non-linearity induces a *sensitive dependence on dynamical equations*. The LSE group then draws the further conclusion that since climate models are (a) heavily idealized and (b) non-linear, we should have low confidence in "decision-relevant" quantitative climate predictions—though they acknowledge that just how sensitive the models are to SME is a question the requires further investigation (Frigg, Bradley, et al. 2014, 48).

The arguments of the LSE group have spawned a series of responses (Goodwin and Winsberg 2016; Nabergall, Navas, and Winsberg 2019; Winsberg 2018; Winsberg and Goodwin 2016) from a group of philosophers and scientists associated with the University of South Florida (who I'll term the "USF group" for parallel's sake). The main contention of the USF group is that the analogy between systems that exhibit SDIC and what I above termed sensitive dependence on dynamical equations cannot be made precise for two reasons.[2] First, the space of dynamical equations is topological but not metrical, meaning that there's no general way to say what it means to be an "arbitrarily close" equation (Winsberg and Goodwin 2016, 14). Second, and for similar reasons, the mathematically well-defined property closest to sensitive dependence on dynamical equations shows only that small dynamical misrepresentations *can* amplify into large errors in the representation

[1] Also worth noting are Mayo-Wilson (2015), L. A. Smith (2007), and Thompson (2013), who explore the possibility of a dynamical analogue of chaos without drawing the same epistemic conclusions vis-a-vis climate science.

[2] They have also advanced a number of other objections, most notably that the central motivating example of the LSE group involves methods that are distinct from those used in much of climate modeling (see, e.g., Winsberg and Goodwin 2016, 12,15). The point is well taken—the example employed by the LSE group is not sufficient to establish general conclusions—but, as we'll see, the relevant epistemic worries have nothing to do with the features specific to that example.

of later states of the target system not that it *will* (Nabergall, Navas, and Winsberg 2019, 11-12). They conclude that there's no general threat to quantitative climate predictions stemming from "infinitesimally small" dynamical misrepresentations (Nabergall, Navas, and Winsberg 2019, 21). Acknowledging, of course, that dynamical misrepresentations do pose epistemic problems in some cases, they argue that the failure of the analogy means that we should resist the LSE group's general conclusions; the epistemic implications for decision-relevant quantitative climate predictions must be evaluated individually (Nabergall, Navas, and Winsberg 2019, 20; Winsberg and Goodwin 2016, 16).

3 Chaos and error

Both the LSE and USF groups appear to consider the (alleged) epistemic problem to be one of *error*.[3] In their central thought experiment, for instance, the LSE group present the problem associated with SME as one of erroneous probabilistic predictions: the agent facing SME-related problems "regards events that do not happen as very likely, while he regards what actually happens as very unlikely" (Frigg, Bradley, et al. 2014, 39). Similarly, in their discussion of the primary motivating case study—a project involving generating "decision-relevant" probabilistic predictions about the future climate in Great Britain—they worry that:

> Trying to predict the true climate with structurally wrong models is like trying to predict the trajectory of Mercury with Newtonian models. These models will invariably make misleading (and likely maladaptive) projections beyond some lead time, and these errors cannot be removed by adding a linear discrepancy term derived [solely] from other Newtonian models. (Frigg, L. A. Smith, and Stainforth 2015, 3997)

And the USF group is no different. They echo the language of the LSE group in their own discussion of the motivating example (Goodwin and Winsberg 2016, 1125), and more recently, they've stated that "only strong versions [of chaos] are usually taken to have strong epistemological consequences, since they are *likely* to produce error" (Nabergall, Navas, and Winsberg 2019, 7, note 13).

[3]That said, in their more careful moments, at least, the LSE group can be read as primarily concerned with particular sorts of tradeoffs between precision and certainty (see, e.g., Frigg, Bradley, et al. 2014, 50). In light of the arguments presented in this section, I think that the most charitable interpretation is likely to emphasize this concern over the concern for error. See section 5 for more details.

To be sure, in the context of the examples employed by the LSE group, introducing chaos *while holding the predictions of the agent fixed at a given level of precision* does increase the probability of error. Introducing sensitive dependence to dynamical equations has the same effect. The USF group are also right that the effect is only significant for stronger versions of chaos and (we could add) only significant if the time frames line up in the right way. But these facts don't imply that there's a connection between chaos (or chaos-like behavior) and error *in general*. If there were such a connection, the consequence would be that we can't make accurate and/or precise predictions about chaotic systems—or (more weakly) that the behavior of such systems are generally harder to predict than that of non-chaotic systems. But this simply isn't the case.

The LSE group's own analogy illustrates the point nicely. The errors in Newtonian predictions of the trajectory of Mercury are on the order of mere arcseconds per century—that is, a prediction of where Mercury will appear in the sky a hundred years out will exhibit an error roughly $1/40^{\text{th}}$ the apparent width of the moon. It's hard to argue that such small errors are genuinely maladaptive. And the solar system as a whole is chaotic: eventually—that is, approximately five million years from now—small differences between present conditions will have grown exponentially larger.[4] And yet we're nevertheless able to make astoundingly precise (and "decision-relevant"!) predictions about the locations of various stellar bodies, for the simple reason that the five-million year timescale is totally irrelevant for predictions in the here-and-now.[5] While there are cases in which chaotic behavior creates genuine problems for predictive accuracy, in other words, it's simply illegitimate to draw inferences from either chaotic behavior or the lack thereof to the the existence of such problems without further information.

The same conclusion is suggested by close attention to more precise definitions of chaotic behavior. Consider the common definition of SDIC in terms of Lyapunov exponents. Suppose that there is a system characterized by a state space X and dynamical mapping $f : X \to X$ such that $x_t = f(x_{t-1})$. This system exhibits SDIC if and only if for all y "arbitrarily" close to x_0,

$$d(x_t, y_t) > e^{\lambda t} d(x_0, y)$$

[4]That is, the "Lyapunov time" of the solar system—the time it takes for distances to grow by a factor of e—is approximately five million years.

[5]Well, not totally irrelevant, because the same physical properties that engender the chaotic behavior of the solar system generate attractors that can affect satellite trajectories; see Wilhelm (2019).

where λ, the "Lyapunov exponent," is positive. Essentially: trajectories that are currently "nearby" will grow exponentially farther apart. If we interpret $d(x_0, y)$ as the present uncertainty, then SDIC entails that uncertainty will grow exponentially with time. It's a significant step from uncertainty growing exponentially with time to either a high probability of error at a given time or some sort of guarantee of inaccuracy. To make this step there needs to be a tight relationship between the relevant timescales and (as we saw above) there's no guarantee that the timescale relevant to our predictions will be the same one that's relevant to chaos. Similar comments apply to other technical definitions of chaos. Werndl (2009), for example, shows that a system is "topologically-mixing" if and only if

$$\lim_{n\to\infty} Pr(x_0|x_{-n}) - Pr(x_0) = 0$$

which, in English, says that the probablistic relevance of past events to future events eventually approaches zero. Chaotic systems "lose" information over time, but the mere fact that information is guaranteed to be lost *eventually* doesn't implicate our ability to make precise or accurate predictions now.

The explanation for the disconnect between chaotic behavior on the one hand and predictive inaccuracy on the other is that SDIC defines a notion of physical sensitivity that is *independent of human interests*. Until we specify a timeframe and desired level of precision for a hypothesis, we cannot know what implications SDIC will have for said hypothesis. Since our interests don't map onto physical distances in any consistent way—a few centimeters of error is a disaster in a surgical setting but incredible in astrophysics—SDIC doesn't have any general implications for either error or the probability of error. Similarly, we should expect that the failure of a system to exhibit SDIC—or an SDIC-like property—also has no general implications for error. The contested claims about the analogy between SDIC and sensitive dependence on dynamical equations therefore has no clear or direct implications for the epistemology of climate modeling; like the solar system, climate models could exhibit exponential growth in the distances between alternative trajectories over time frames on the scale of millions of years. Or they could fail to exhibit any growth in distances between trajectories but the starting uncertainty could be too substantial to license "decision-relevant" predictions. Insofar as our concern is something like the probability of error in general,

chaos and chaos-like behavior simply aren't relevant.

4 Failures of safety

To determine whether the presence of dynamical misrepresentations renders "decision-relevant" quantitative climate predictions untrustworthy, we need a different, interest-relative, concept of sensitivity to small errors. My view is that the relevant concept is given by the failure of a kind of safety condition.

Speaking abstractly, when we're concerned with whether we should believe some hypothesis, one relevant desideratum is that the justification for the hypothesis should be *safe*: the degree of support for the hypothesis should be (nearly) the same given nearby alternative background assumptions, where a background assumption is "nearby" to the extent that it has a relatively high probability on the total evidence available.[6] So, for instance: if my evidence for the fact that it is freezing outside is the reading of my thermometer, then the hypothesis is safer in the situation where the thermometer reads -5°C than when it reads -1°C; the former allows for more leeway in the background assumptions concerning the accuracy of the thermometer. When the evidence for a hypothesis rests either fully or partially on a model, the hypothesis is going to more or less safe to the extent that sufficiently small changes to the assumptions of the model don't (substantially) affect the results or outputs of the model. The reasoning here is the same. If the hypothesis is only supported by the model given precise and risky assumptions, then there's a relatively high chance that these assumptions don't hold. By contrast, if the hypothesis is supported regardless of whether we use the specific assumptions in question or any one of a number of nearby assumptions, then the hypothesis is safe.

Intuitively, safety is going to be related to the probability of error at least under conditions in which there's some degree of uncertainty about the quality of the evidence. Since humans are not ideal reasoners, we're often in situations in which we don't know how likely some hypothesis is on our evidence. So, for instance, we might know that we should be confident in P if Q is true, but not

[6]This notion of safety is essentially the one found in Reed (2000) and Staley (2004), and is tightly connected to G. E. Smith (2002, 2014)'s discussion of "*quam proxime*" reasoning. Like other safety conditions, the best way to define this one precisely is in terms of possible worlds and a distance measure between them, though what we want is a graded measure that allows for higher and lower degrees of safety. I take it that how this all works intuitively is clear enough for present purposes.

either whether Q is true or what our confidence in P should be given relatively likely alternatives to Q.[7] Why might we be in this situation? One common and relevant reason is that our evidence relies on an idealized model; since the inner workings of models are often "opaque" (Humphreys 2004), we can't know *a priori* whether or not the idealizations in question merely serve to simplify the problem in a harmless manner or, by contrast, whether they substantially affect the output of the model.[8] In other words, we don't know whether P is *safe*, whether it would still be justified given small changes to the background assumptions. If it is safe, then the evidence is trustworthy and provides good reason to believe that P; if it isn't, then the total evidence does not provide reason to believe that P. If we accept P, therefore, safety and error will be inversely correlated: the safer the hypothesis, the higher its overall justification, and thus the lower the chance that it is has been accepted in error.

The foregoing is highly abstracted from the practices of science. Consider, therefore, the derivation of inverse-square gravity from Kepler's first law.[9] Suppose that Kepler's first law holds *exactly*, meaning that sun is at the focus of each planet's elliptical orbit and that the distance function between planet and sun is

$$d = A\frac{(1-\epsilon^2)}{1-\epsilon\cos\theta}$$

where A is the long arm of the ellipse and ϵ the eccentricity. In combination with some other information about the nature of ellipses, this equation entails that the acceleration of the planet is proportional to the inverse of square of the distance ($a \propto d^{-2}$). It's thus possible to derive the inverse-square law from Kepler's first law. In the context of the present discussion, however, this derivation faces two problems. First, there was little evidence available that Kepler's first law held *precisely* (and, in fact, it doesn't): the difference between an ellipse with the sun at a focus and

[7] Epistemologists term cases like these instances of "higher-order uncertainty." There's disagreement concerning whether we can *rationally* have higher-order uncertainty (see, e.g., Dorst 2019; Titlebaum 2015); I won't take a stance on that here. My concern is that as non-ideal agents we frequently are uncertain when we would rather not be.

[8] Isn't this a case in which we know Q to be false? No, though defending this point adequately here would take us too far afield. The central idea is that it's a mistake to read Q as a claim about the truth of the idealized model rather than as a claim about its (non-literal) accuracy (Frigg and Nguyen 2019), adequacy-for-purpose (Parker 2009), or reliability (Dethier 2019).

[9] I'm borrowing this example from G. E. Smith (2002). Smith's point, which is worth emphasizing, is that the safety failure present in this example provides the best explanation for why Newton himself didn't derive the inverse-square law in this manner, preferring instead the evidence provided by the apsides of the planets (Newton 1727/1999, 802), relative to which the hypothesis is *extremely* safe.

an ellipse with the sun at the center is virtually undetectable with 17$^{\text{th}}$ century tools.[10] The first problem with the derivation, then, is that we're uncertain whether the assumptions built into it hold precisely.

The second problem is that the derivation is extremely sensitive to small deviations from Kepler's first law. As just noted, at low eccentricity, there's very little difference between an ellipse with the sun at the focus and one with the sun at the center. The distance function for the latter is given by

$$d = A\sqrt{1 - \epsilon^2 \sin^2 \theta}$$

And this function, in combination the same assumptions about the nature of ellipses, entails that the acceleration of the planet is proportional to the distance *directly* ($a \propto d$). The derivation from Kepler's first law therefore provides extremely poor evidence for the conclusion in the sense that it relies on a particular assumption holding precisely when the best evidence available only indicates that the assumption holds approximately.

This case provides an exemplar of a safety failure in a number of respects. Recall: safety failures arise because the quality of the evidence varies dramatically with small changes in background assumptions. Here the changes to background assumptions are small not because the two equations are nearby in any mathematical sense but because the evidence makes both assumptions relatively likely. And the difference in the quality of the evidence is dramatic because of our particular choice of how to divide up the hypothesis space: what matters is that d and d^{-2} make for extremely different theories of gravitation. If our hypothesis was simply that there is *some* relationship between distance and acceleration, there would not be a safety failure to be found. It is also exemplary with regards to effects: the safety failure makes it likely *on our evidence* that if we accept the hypothesis, we're going to do so erroneously—which is just to say that when the hypothesis fails to be safe, the evidence doesn't give us much reason to believe it. (Though, of course, and as evidenced by this example, other evidence might; see note 9.)

The definition of safety given in this section provides a notion of sensitivity that is appropriately

[10]Kepler argued for his first law by showing that it held to a high approximation with respect to Mars (Miyake 2015); taking it to hold to a high approximation with respect to the other planets is a relatively risky inductive move—and one that Newton knew his own theory would show to be invalid (Newton 1727/1999, 817-18).

dependent on human interests. The class of objects that there is sensitivity to are defined or identified according to our epistemic abilities: in the modeling context, it's the class of assumptions that are empirically adequate by our standards. Similarly, whether or not there is sensitivity to the differences between these representations depends on our interests and concerns insofar as those affect how precise we want or need our hypotheses to be. There's a *rough* analogy to SDIC or chaos here in that a safety failure involves a "growth" in "distances" in an interest-dependent sense: the initial distance is small relative to our ability to distinguish between different scenarios and the latter one large relative to our desire for precision. But this is connection is not mathematically precise. In particular, safety failures are not analogous to SDIC in the ways that the USF group argues present problems for the LSE group.[11] There's no interest-independent distance measure to be placed on either the different starting characterizations of the system or the resulting equations. The different distance equations are similar just in the sense that they're both empirically adequate in the given situation; the different relationships between acceleration and distance are dissimilar in the sense that their broader fit in the theory is dramatically different. Further, we haven't shown that *any* nearby deviation from Kepler's laws (or even Kepler's first law) will lead to an arbitrarily different relation between acceleration and distance. All that we've shown that there is a particularly salient alternative that has this effect.

This section has provided an appropriate notion of sensitivity to employ in getting clearer about the debate over chaos. In the next section, I'll argue for a reinterpretation of the LSE group's arguments in terms of safety failures.

5 Reinterpreting the LSE group

I think that the arguments presented by the LSE group are important, but they don't show that the possibility of small dynamical errors should cause us to lower our confidence in various claims supported by climate models. Instead, they should be interpreted as offering an explanation of (empirically-ascertained) levels of model precision in terms of small dynamical errors—an explanation that, if true, has important implications for which projects in climate science are likely to be successful.

[11] There are other disanalogies as well, besides those at issue in the debate surveyed in §2. For instance, neither initial conditions nor time has any role in this case, though both are essential to the understanding of SDIC.

The main motivation behind this interpretation of the LSE group is that the arguments that they offer are neither necessary nor sufficient to establish that we should be less confident in the claims supported by climate models in general. They are not sufficient because they would need to show that climate scientists have generally been overconfident in modeling results—but climate scientists are well aware that climate models can be highly misleading, even in the aggregate (Knutti, Furrer, et al. 2010). They are not necessary because general considerations about safety failures provide much less powerful (and precise) evidence for caution about specific climate hypotheses than is provided by the empirical evaluation of climate models. Evaluation studies provide evidence not just about the degree of confidence licensed by a given model, but also about where the models excel, where they struggle, and what assumptions account for these struggles. Any general considerations about safety failures are likely to simply be swamped by the empirical evidence from this domain.

Of course, the LSE group is well aware of this empirical literature—as evidenced by their prior work drawing out the implications of it for decision-making (Oreskes, Stainforth, and L. A. Smith 2010; Stainforth et al. 2007). My suggestion is that we should read their arguments concerning chaos through the lens of this earlier work.[12] Specifically, we should view the combination of small dynamical errors and system complexity as providing an explanation for why climate models are only able to achieve certain levels of precision and accuracy. In giving this explanation, the LSE group is stressing that our inability to draw conclusions about local policy from climate models isn't a temporary defect of these models. On the contrary, hypotheses about how climate change is going to affect a town, region, or (small) country are simply too sensitive to small changes in modeling assumptions, and we're not likely to reach a point any time in the near future where we have the ability to determine which of these assumptions are true. That's just the nature of the system (a conclusion, I'll note, that is widely shared among climate scientists; see Knutti and Sedlàček 2013). In other words: hypotheses that we know are unsafe based on our empirical evaluations of the models are likely to remain unsafe—and thus, as the LSE group explicitly suggests, we need methods for determining how to make decisions under conditions where the quality of our evidence is uncertain in precisely this way.

If this is the correct interpretation *and* the arguments given in prior sections are correct, then the way that the LSE group has presented their arguments for this conclusion is misleading; the

[12]Here I'm following Mike Goldsby and Greg Lusk.

analogy to SDIC is largely irrelevant to whether or not the present and future levels of uncertainty about the dynamics are likely to undermine the evidence for future climate hypotheses. My experience is that they're doubly misleading to those unfamiliar with precise definitions of chaos—many of those presented with the arguments seem to automatically assume that chaotic behavior means that *anything* goes. Many of the USF group's criticisms (particularly those not discussed above) are aimed directly at this point: the presentation and rhetoric of the analogy to chaos, they contend, doesn't align with the more limited conclusions that the LSE group wants to draw (see Goodwin and Winsberg 2016; Winsberg and Goodwin 2016). As we've just seen, however, whatever the disconnect between the rhetoric and the arguments in the LSE group's papers, there isn't a genuine worry that their arguments might—even if successful—undermine much more than they intend: on this interpretation, the arguments simply don't motivate changing our confidence in any particular results of the models; they motivate "only" changing our confidence that we'll be able to get well-justified decision-relevant predictions out of the models any time soon.

To be clear, I am not arguing that the argument just sketched is correct. Nevertheless, the conclusion is interesting and the arguments itself has the advantages of fitting nicely with the prior work of the LSE group, not relying on mistakes concerning the relationship between chaos and error, and not—if successful—implicating far more of climate science than can be plausibly be justified on the evidence appealed to. We thus have good reason to interpret the LSE group in this manner, even if the argument ends up being unsound.

6 Conclusion

In this paper, I've argued that the details of dynamical analogues of chaos are largely irrelevant to the epistemological questions raised in the recent debate over them. Mathematically interesting as the alleged analogy may be, a tight analogy to chaos is neither necessary nor sufficient for the kinds of epistemic error that motivate the debate. The type of *epistemic* sensitivity relevant to error is better captured by failure of a kind of safety condition: what's worrying about dynamical misrepresentations is that they undermine the evidence provided by the model. Once the irrelevance of chaos is recognized, it becomes clear that the upshot of the debate is not whether models are likely to be erroneous but an explanation for why models are not more precise than they in fact

are.

References

Dethier, Corey (2019). How to Do Things with Theory: The Instrumental Role of Auxiliary Hypotheses in Testing. *Erkenntnis* (online first).

Dorst, Kevin (2019). Higher-Order Uncertainty. In: *Higher-Order Evidence: New Essays*. Ed. by Mattias Skipper and Asbjørn Steglich-Petersen. Oxford: Oxford University Press: 35–61.

Frigg, Roman, Seamus Bradley, et al. (2014). Laplace's Demon and the Adventures of His Apprentices. *The Journal of Philosophy* 81.1: 31–59.

Frigg, Roman and James Nguyen (2019). Mirrors Without Warnings. *Synthese* (online first).

Frigg, Roman, Leonard A. Smith, and David A. Stainforth (2013). The Myopia of Imperfect Climate Models: The Case of UKCP09. *Philosophy of Science* 80.5: 886–97.

— (2015). An Assessment of the Foundational Assumptions in High-Resolution Climate Projections: The Case of UKCP09. *Synthese* 192.12: 3919–4008.

Goodwin, William Marc and Eric Winsberg (2016). Missing the Forest for the Fish: How Much Does the 'Hawkmoth Effect' Threaten the Viability of Climate Projections? *Philosophy of Science* 83.5: 1122–32.

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

Knutti, Reto, Reinhard Furrer, et al. (2010). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate* 25.10: 2739–58.

Knutti, Reto and Jan Sedlàček (2013). Robustness and Uncertainties in the New CMIP5 Climate Model Projections. *Nature Climate Change* 3: 369–73.

Mayo-Wilson, Conor (2015). Structural Chaos. *Philosophy of Science* 82.5: 1236–47.

Miyake, Teru (2015). Underdetermination and Decomposition in Kepler's *Astronomia Nova*. *Studies in the History and Philosophy of Science Part A* 50: 20–27.

Nabergall, Lukas, Alejandro Navas, and Eric Winsberg (2019). An Antidote for Hawkmoths: On the Prevalence of Structural Chaos in Non-linear Modeling. *European Journal for Philosophy of Science* 9.21: 1–28.

Newton, Isaac (1727/1999). *Mathematical Principles of Natural Philosophy*. Trans. by I. Bernard Cohen and Anne Whitman. 3rd edition. Berkeley: University of California Press.

Oreskes, Naomi, David A. Stainforth, and Leonard A. Smith (2010). Adaptation to Global Warming: Do Climate Models Tell Us What We Need to Know? *Philosophy of Science* 77.5: 1012–28.

Parker, Wendy S. (2009). Confirmation and Adequacy-for-Purpose in Climate Modelling. *Aristotelian Society Supplementary Volume* 83.1: 233–49.

Reed, Baron (2000). Accidental Truth and Accidental Justification. *The Philosophical Quarterly* 50.198: 57–67.

Smith, George E. (2002). From the Phenomenon of the Ellipse to an Inverse-Square Force: Why Not? In: *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics*. Ed. by David Malament. La Salle: Open Court: 31–70.

— (2014). Closing the Loop: Testing Newtonian Gravity, Then and Now. In: *Newton and Empiricism*. Ed. by Zvi Beiner and Eric Schliesser. Oxford: Oxford University Press: 262–351.

Smith, Leonard A. (2007). *Chaos: A Very Short Introduction*. Oxford: Oxford University Press.

Stainforth, David A. et al. (2007). Confidence, Uncertainty and Decision-Support Relevance in Climate Predictions. *Philosophical Transactions of the Royal Society Series A* 365.1857: 2145–61.

Staley, Kent W. (2004). Robust Evidence and Secure Evidence Claims. *Philosophy of Science* 71.4: 467–88.

Thompson, Erica (2013). Modelling North Atlantic Storms in a Changing Climate. PhD dissertation. Imperial College, London.

Titlebaum, Michael (2015). Rationality's Fixed Point (or: In Defense of Right Reason). *Oxford Studies in Epistemology* 5: 253–294.

Werndl, Charlotte (2009). What Are the New Implications of Chaos for Unpredictability? *The British Journal for the Philosophy of Science* 60.1: 195–20.

Wilhelm, Isaac (2019). Celestial Chaos: The New Logics of Theory-Testing in Orbital Dynamics. *Studies in History and Philosophy of Science Part B* 65: 97–102.

Winsberg, Eric (2018). *Philosophy and Climate Science*. Cambridge: Cambridge University Press.

Winsberg, Eric and William Marc Goodwin (2016). The Adventures of Climate Science in the Sweet Land of Idle Arguments. *Studies in History and Philosophy of Science Part B* 54: 9–17.

Wishful Intelligibility, Black Boxes, and Epidemiological Explanation

Marina DiMarco
Department of History & Philosophy of Science
University of Pittsburgh

Abstract

Epidemiological explanation often has a "black box" character, meaning the intermediate steps between cause and effect are unknown. Filling in black boxes is thought to improve causal inferences by making them intelligible. I argue that adding information about intermediate causes to a black box explanation is an unreliable guide to pragmatic intelligibility because it may mislead us about the stability of a cause. I diagnose a problem that I call wishful intelligibility, which occurs when scientists misjudge the limitations of certain features of an explanation. Wishful intelligibility gives us a new reason to prefer black box explanations in some contexts.

Acknowledgments:

I am grateful to Jim Woodward, Anya Plutynski, Mike Dietrich, Kareem Khalifa, Jonathan Fuller, Kathleen Creel, Dasha Pruss, the Pitt HPS Works in Progress community, and participants at POBAMz 2020 for generous and thoughtful discussions about this argument.

1 Introduction

Epidemiological explanation often has a "black box" character, meaning the intermediate steps between some putative cause and effect of interest are unknown. The black boxes of epidemiological explanation have been variously described as mere predictive heuristics; as obstacles, or even threats to scientific understanding. Philosophers and epidemiologists alike have argued that specifying intermediate causes to fill in black boxes improves causal explanations by making them intelligible, and better targets for public health intervention (Machamer, Darden, and Craver 2000; Hiatt 2004; Russo and Williamson 2007). Specifying the links between the ends of a causal chain is supposed to confer certainty, understanding, and reasons to expect a causal relationship to be stable or invariant across populations of interest.

I argue that adding information about intermediate causes can be an unreliable guide to improving epidemiological explanation because it may mislead us about the stability of a causal relationship and may convey a false sense of understanding. I diagnose this as an instance of a more general problem that I call wishful intelligibility, which occurs when scientists misjudge the limits of the pragmatic benefit conferred by certain features of an explanation. To illustrate this, I consider an example of epidemiological explanation involving the social determinants of health. My argument offers a new reason to prefer black box explanations in some contexts: not despite, but because of, their lack of information about intermediate causes. This preference has the consequence that filling in black boxes is not a necessary source of intelligibility, but a contingent one.

2 Black Boxes and Intelligibility

Specifying the links between the ends of a causal chain is supposed to improve our understanding of an epidemiological cause, but attempts to account for the intelligibility

produced by filling in black boxes vary considerably. For mechanists like Machamer, Darden, and Craver (2000, 21), filling in intermediate links of a causal chain between some cause C and effect E at the preferred level of detail for a given scientific field makes the relationship between C and E more intelligible. This sense in which filling in black boxes is supposed to confer understanding on this account is rather vague, leaving open the possibility that their "intelligibility" is akin to the phenomenological "sense of understanding" that Trout (2002) convincingly condemns as an unreliable indicator of explanatory goodness. By contrast, other accounts of epidemiological explanation tether the value of filling in black boxes to the goal, broadly speaking, of using epidemiological explanations to design effective public health interventions (Russo and Williamson 2007; Illari 2011; Broadbent 2011). Following de Regt (2017), I'll call this pragmatic intelligibility. I will focus on the pragmatic intelligibility argument, both because it seems to evade the phenomenological critique and because it bears a more obvious relationship to the design of public health policy.

One way that filling in black boxes is supposed to confer pragmatic intelligibility is by informing our inferences about the stability of epidemiological causes; that is, our expectation that a causal relationship observed in one context will also hold in other contexts of interest. Because many epidemiological causal relationships are observed in population studies, the design of effective public health interventions must often attend to the potential efficacy of such interventions both within and beyond the original conditions in which a causal relationship is observed (Dupré 1984). Russo and Williamson (2007) famously argue that filling in black boxes supports epidemiological causal inference in just this way: that filling in a black box with evidence of a plausible mechanism tells us about the stability of a cause, and supports an expectation that the causal relationship will hold in contexts of interest that differ from the one in which it was observed. This feature would, if true, make such evidence key to the use of epidemiological causes with regard to

the design of public health interventions. In section 3, I will show why this is not necessarily what we ought to expect.

3 Stability and Intermediate Causes

I agree that stability is one of the most important features of epidemiological inference relative to the goal of reliable public health intervention. However, I argue that filling in black boxes can lead us to both under- and overestimate the stability of epidemiological causes. In practice, this makes filling in black boxes an unreliable guide to the pragmatic intelligibility of epidemiological explanation.

Because Russo and Williamson take the increase in stability of a cause to mean that it is expected to hold in conditions different from the ones in which the original experiment or observation took place, I take it that they have in mind something like Woodward's (2010) notion of the stability of causes. On this account, there is a causal relationship between two variables C and E just in case some intervention on the value of C produces a change in the value of E that proceeds only through the change in C (Woodward 2000). Interventionist causal relationships are stable or invariant to the extent that they hold over a more or less universal set of background conditions, making causal stability a matter of degree (Woodward 2000, 2010). Although this account is particularly amenable to epidemiological practice because a Woodward intervention need not be the result of intentional human manipulation, one does not need to be an interventionist to appreciate this notion of stability.

Filling in black boxes entails adding links in a causal chain. As Woodward (2010) points out, however, for any causal chain, the set of background conditions or domain of invariance over which the entire chain is stable is limited to the extent to which the stability of all links in the chain overlaps. This means the chain is limited not only by the

domain of the least stable link, but also by the parts of that domain of invariance that are shared with the other links. Because intermediate causes constrain the stability of the chain in this way, adding information about intermediate links cannot, by itself, increase our confidence in the stability of the entire chain.

3.1 Underestimating Stability

With this account of stability in mind, a focus on filling in black boxes can mislead us about the stability of a causal relationship in at least two ways. First, identifying a single causal chain from C to E may be misleading with respect to the stability of a cause that produces its effect by multiple independent pathways. As Mitchell (2002), Fehr (2004), Dupré (2013) and Howick et al. (2013) argue, specifying a single set of intermediate steps between cause and effect restricts our assessment of the relationship between C and E to one particular causal chain, when in fact there may be several pathways from C to E. For instance, a causal variable like socioeconomic status might cause cancer by way of its effects on stress, nutrition, access to preventive care, and so on. Multiple pathways between a single cause and effect may be stable over different background conditions. This means that filling in a black box with a single causal chain can lead us to underestimate the stability of a causal relationship by confining our expectation to a single, overly narrow domain of invariance.

3.2 Overestimating Stability

As Fehr (2004) points out, our interest in causal intermediates need not commit us to the myopia of *single* mechanistic explanation. However, more sophisticated efforts to fill in black boxes are still subject to a second set of concerns about stability: namely, that filling in black boxes can lead us to overestimate the stability of a causal chain when we overlook the challenges of integrating multiple indirect causes. This is because filling in a black box

is often a process of what Mayo-Wilson (2014) calls "piecemeal causal inference." The intermediate steps between a cause and effect in epidemiology are frequently inferred in independent research contexts. As Baetu (2014), Mayo-Wilson (2014), and others have pointed out, integrating causal variables that are inferred in different research contexts increases underdetermination and uncertainty about causation. This problem is particularly pernicious and complex when it comes to assessing the stability of causal chains in epidemiology.

At least three features of epidemiological practice contribute to this difficulty. First, ethical constraints make it the case that many causal inferences in epidemiology cannot be made on the basis of manipulations or interventions in human populations. This means many links in a putative causal chain are thought to be stable with respect to humans on the basis of extrapolation from animal models, or retrospective analyses of so-called natural experiments. Extrapolation and inferences of external validity are inherently epistemically risky business (cf. Reiss 2019). Second, causal variables within the same chain are often measured and described with very different degrees of precision, and at different spatial and temporal scales in different research contexts. Finally, scientists in different research contexts often measure causal variables with respect to different background conditions of interest. For instance, social epidemiologists (e.g. Krieger 2008) are especially concerned to include possible social determinants of health, like socioeconomic status, as variables in their analyses, but other researchers interested in intermediate causes of the same effects, like epigeneticists, may not measure the socioeconomic status of their subjects at all. Even similar variables described and measured at similar scales are not consistently accounted for across studies in the same field. For example, "neighborhood" is variously measured by zip code, census tract, or county (Shavers 2007). When researchers want to integrate causal inferences to fill in a black box between, for instance, neighborhood and cancer mortality, these differences limit the extent to which piecemeal causal inference can tell us a causal

relationship is stable with regard to the values of these differently described variables. These failures to consistently define and co-measure background condition variables are a problem because they mean that inferences about whether links are stable with regard to the same background conditions may be much less certain than they appear.

When a black box causal relationship is filled in using intermediate causal inferences assembled from diverse research contexts, it presents a unique challenge for assessing the stability of the original causal relationship of interest. Because links in the same causal chain are often inconsistently described, measured at different scales, and demonstrated with respect to different background conditions, it is often impossible (or at least intractable) to assess the extent to which multiple links in a causal chain share a domain of invariance at all. Integrative inferences about the stability of the entire chain become much more complex. At a minimum, these factors make it difficult to identify a lowest common denominator, or least stable link in a causal chain. Failure to attend to these features of piecemeal causal inference can lead us to overestimate the stability of a causal relationship or to make an inference about its stability that is not justified by the available evidence.

3.3 Wishful Intelligibility

Since filling in black boxes is at best an unreliable guide to stability, and stability is critical to the goal of designing epidemiological interventions, it follows that we should not expect filling in black boxes to confer pragmatic benefit to epidemiological causes by way of improving our inferences about stability in all contexts. Instead, this assumption may lead us to be inappropriately confident in our understanding of a cause and in our estimation of stability in particular. We should not expect filling in black boxes to be conducive to the goals of epidemiological inference in cases where such goals depend on information about the stability of a cause; this is a contingent, rather than a necessary, source of pragmatic intelligibility.

To mischaracterize the intelligibility of a causal claim is to misjudge the extent to which we understand it. This may (mis)inform the design of interventions, with serious consequences for public health policy. Unjustified attributions of pragmatic intelligibility constitute a special kind of second-order misunderstanding; namely, a failure to correctly assess the extent of the pragmatic benefit of some particular feature of an explanation (Steel 2016; cf. Trout 2002). I call this wishful intelligibility. An unqualified preference for filling in black boxes in pursuit of stability can lead us to wishful intelligibility (for a similar argument, see Broadbent 2011).

Wishful intelligibility has an obvious affinity with the more general problem of wishful thinking in the literature on science and values. Broadly speaking, wishful thinking may occur when certain values or cognitive biases lead us to form an otherwise unjustified or ill-justified belief; these biases may include but are not limited to a desire for the belief to be true (see Anderson 2004; Steel 2018). By contrast, wishful intelligibility concerns not whether a belief or claim is justified in general, but rather, whether we have good reason to expect that some feature of an explanation is conducive to its use in a specific context. That is, it concerns a particular set of beliefs: those about the pragmatic intelligibility conferred by certain features of an explanation.

4 Multilevel Causes of Cancer

In section 3, I argued that filling in black boxes with intermediate causes can be misleading with respect to the stability of a causal relationship, and that filling in black boxes can be conducive to wishful intelligibility with regard to epidemiological explanation. Gehlert and colleagues' (2008) multilevel model of the social environment as a cause of cancer, from the University of Chicago's Center for Interdisciplinary Health Disparities Research (CIHDR), shows the difficulty of estimating the stability of a piecemeal causal inference. Their work

is particularly interesting because it purports to have implications for the design of future public health interventions.

4.1 The Social Environment and Breast Cancer

This multilevel model fills in links in a (putative) causal chain by assembling multiple local causal inferences from separate studies. A CIHDR study evaluates the social environment of black women newly diagnosed with breast cancer in "predominantly black neighborhoods of Chicago," using interviews and "publicly available data geocoded to the women's addresses" (2008, 343). Gehlert et al. (2008, 344) argue that features of the built environment, such as dilapidated housing and crime, contribute to social isolation. One example cited, Sampson et al. (1997) is an investigation of the effects of "collective efficacy" (defined as neighborhood social cohesion) on violence in Chicago in 1995.

The next steps in the chain are the sequential links between social isolation, psychological states, and stress hormone responses. Here Gehlert et al. appeal to a combination of human and rodent studies to argue that social isolation affects HPA axis regulation and glucocorticoid (stress hormone) signaling via epigenetics. Since glucocorticoid levels have been linked to suppressed immune function elsewhere in the literature, the authors conclude that stress hormone regulation constitutes a means by which social isolation can "get under the skin" to cause cancer and promote tumor cell survival (2008, 343).

This model is a clear case where piecemeal assembly of links between "levels" does not (by itself) support any inference about the stability of the whole chain. Importantly, each step is inferred with respect to different background variables, and few, if any, of the same background conditions are measured for any two links in the chain. Some links, like the association between social isolation and mammary gland tumors, are manipulated in a laboratory environment, using model organisms, while others are observed in humans.

Qualitatively similar links in the chain are differently described: Sampson and colleagues' study of collective efficacy is performed in the same Chicago neighborhoods as the CIHDR study of newly diagnosed cancer patients, but using a notion of "neighborhood clusters" to combine 847 census tracts into 343 clusters, from more than a decade earlier. Setting aside the fact that these neighborhoods have probably changed over ten years, the extent to which these neighborhood clusters overlap spatially with the geocoded data used to measure neighborhoods in the CIHDR project is unclear. Some links, such as epigenetic regulation of altered HPA axis-related gene expression, are apparently assumed to be stable across human populations over which they may not have been measured at all. Gehlert and colleagues borrow links from several research contexts to fill in the black box between neighborhood and tumorigenesis.

Filling in the black box between the social environment and cancer incidence does not justify an expectation that the relationship between them will be stable. It does not even tell us anything reliable about how stable we might expect it to be. Instead, it may lead us to attribute wishful intelligibility to a social epidemiological cause where it is lacking, and to misrepresent the limits of (and risks entailed by) this explanation as a potential basis for public health policy.

4.2 Integration, Stability, and Translation

One might object that the issue in this and similar cases is that we really ought to have something stronger (or at least better confirmed) in mind with regard to filling in black boxes (see Illari 2011). However, attention to the specific features of integration that might justify estimations of stability in such cases is notoriously absent from the Russo-Williamson account and from other arguments for filling in black boxes (Illari 2011; Plutynski 2018). We can have good evidence for each link in a causal chain and yet have relatively little evidence that these links are stable over some shared set of relevant

background conditions (see Broadbent 2011). It is much more difficult to integrate evidence from diverse contexts to show that each link in a chain is stable with regard to the same conditions of interest than it is to justify inferences about individual links in a chain. Thus, assessing the stability of a causal chain can come apart from (and be more burdensome than) merely providing evidence of a causal chain, or providing evidence of individual links.

Importantly, this does not mean that we ought to be pessimistic about multilevel causal models or about the so-called 'social determinants of health' more broadly. Rather, it presents an opportunity to negotiate what does make for better and worse attributions of stability. At a minimum, we might expect that coordination among researchers to improve standardization and co-measurement of causal variables and background conditions, together with explicit evaluation and justification of background assumptions across a causal chain, would improve integration and inferences about the stability of a causal relationship. Many translational epidemiologists have recently turned their attention to these and other related features of knowledge integration in epidemiology with the goal of improving and expediting the translation of biomedical research into public health interventions (e.g. Ioannidis et al. 2013).

Following O'Malley and Stotz (2011) and others, I take it that the details of successful knowledge integration in epidemiology will be largely pragmatic and contextually specific, and most importantly that they will admit of degrees. My main concern is that the wishful intelligibility of filling in black boxes leads us to overlook these considerations entirely. In section 5, I argue that black boxes may prevent this sort of oversight in some contexts by preserving epistemic humility about epidemiological causes.

5 Preferring Black Boxes

So far, I have focused on the extent to which filling in black boxes may confer pragmatic intelligibility, by way of stability. Now, I will consider the relationship between intelligibility and black boxes themselves.

Previous endorsements of black boxed explanations in epidemiology have focused on the utility of black boxes in designing interventions despite ignorance of intermediate links in a causal chain (Cranor 2017; Plutynski 2018). These accounts often appeal to some version of the argument from inductive risk, on which, broadly speaking, the ethical consequences of error play a normative role in determining the amount and kind of evidence necessary to justify accepting or rejecting a hypothesis, and, by extension, the decision to intervene in a particular way (Douglas 2000; Steel 2016). They purport to justify a preference for (or more accurately, a tolerance of) black boxes when the costs of agnosticism or inaction outweigh the possible negative consequences of ignorance about the intermediate steps between some putative C and E, or when the costs of providing evidence for a plausible set of intermediate steps are too high, given the projected benefit of additional detail.

While I am sympathetic to these accounts, I am concerned to show that, by the same token of inductive risk, there are circumstances in which we ought to prefer black boxed explanations not despite, but because of their lack of information about causal intermediates. This is because black boxes can prevent wishful intelligibility, especially where stability is concerned. Imagine a case in which misjudging the stability of a cause can be expected to have serious consequences for the success of some possible intervention. Since filling in a black box can actively mislead us with respect to the stability of a cause, we may prefer a black box for purposes of designing such an intervention. Wishful intelligibility often has a price.

The above discussion of stability can help to predict the contexts in which we might be

particularly susceptible to these mistakes. In cases where we have reason to expect a complex or nonspecific, multi-causal structure, we might be misled by estimating stability from a single causal chain. Further, I have argued that piecemeal causal inference (and especially multilevel piecemeal causal inference) may lead us to overestimate the stability of a causal chain, especially when links in the chain are poorly integrated. In both cases, black boxed explanations capture and convey an appropriate sense of uncertainty about the stability of some cause. In such cases, agnosticism about the stability of a cause may be more conducive to the design of effective interventions than wishful intelligibility would be. In this sense, my argument is concordant with Trout's (2002, 212) condemnation of the risks of "counterfeit understanding." Admittedly, the value of black boxes in these cases is indexed to the costs of being wrong about stability. However, given that stability is lauded as a critical determinant of which epidemiological causes make for good interventions, I take it that these concerns are relevant to a non-trivial number of cases.

Black boxes may thus preserve a certain humility, akin to what Pickersgill (2016) calls "epistemic modesty" about the stability of a cause that is conducive to the design of good interventions and the avoidance of bad ones. This means that black boxes are not merely to be preferred *despite* the risk of unknown intermediates. Rather, a black box is preferable to evidence of a causal chain in cases where the consequences of error about stability are sufficiently undesirable. When we have thorough and well-integrated knowledge of a causal structure, my concerns about under- and overestimating stability are less troubling. Given that the appropriateness of black boxes (and of filling in) is contextual and specific, these considerations put black boxes on par with evidence of intermediate linking causes as features of explanations that may contribute to their pragmatic intelligibility in a contextual and contingent manner.

Of course, this does not mean that black boxes are to be preferred to information about intermediate causes in all or even most epidemiological explanations. The problem lies not

with information about intermediate causes, but with the inferences we are tempted to make about stability on the basis of this information. Furthermore, I have by no means exhausted the arguments for filling in black boxes in epidemiological explanation, and I expect we may have good reasons to prefer filling-in that outweigh these concerns about stability in some contexts. Rather, my argument shows that black boxes deserve the same status of contextual relevance to pragmatic intelligibility as do other features of epidemiological explanations, not despite, but because of, the absence of information about intermediate causes. This means that their inclusion (or filling-in) should be a matter of transparent negotiation and justification, rather than a default preference one way or the other.

6 Conclusions

Wishful intelligibility is a helpful diagnosis for the mistaken expectation that filling in black boxes is a good guide to causal stability. Filling in black boxes is by no means the only possible source of wishful intelligibility in epidemiology or elsewhere, nor does it always have this effect. Rather, I have been concerned to argue about specific epidemiological cases precisely because the features of an explanation that make it conducive to some particular goal are contextual and specific. Similarly, I have shown that there is a positive role for black boxes in preventing wishful intelligibility and in preserving epistemic humility about the stability of complex causes in some cases of interest to epidemiologists.

These cases seem to have something in common; namely, incomplete, limited, or poorly integrated knowledge of a complex causal structure. We might worry that my account could mistakenly preclude a positive and epistemically responsible role(s) for various departures from the whole truth (à la Elgin 2007) or that a preference for black boxes in such contexts perpetuates a harmful role for "ideal science" which may cripple important

regulation and delay helpful interventions (Cranor 2017). But such a preference does not depend on a false dichotomy between ideal science and wishful intelligibility; after all, black boxes are themselves departures from the whole truth. Incomplete or poorly integrated information about a complex causal structure need not paralyze the design of public health interventions; it merely makes our assessments of stability a matter of second-order inductive risk. My account invites transparency and justification for such decisions on a case-by-case basis, and recommends specific ways in which the state of current epidemiological knowledge should inform these considerations. To the extent that these measures avoid wishful intelligibility, they make for more trustworthy reasons to intervene in a particular way, not less.

References

Anderson, Elizabeth. 2004. "Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce." *Hypatia* 19 (1): 1–24.

Baetu, Tudor M. 2014. "Models and the Mosaic of Scientific Knowledge. The Case of Immunology." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 45 (March): 49–56.

Broadbent, Alex. 2011. "Inferring Causation in Epidemiology: Mechanisms, Black Boxes, and Contrasts." In *Causation in the Sciences*, edited by Phyllis McKay Illari, Federica Russo, and Jon Williamson, 45–69. Oxford: Oxford University Press.

Cranor, Carl F. 2017. "How Demands for Ideal Science Undermine the Public's Health." In *Tragic Failures: How and Why We Are Harmed by Toxic Chemicals*. Oxford: Oxford University Press.

de Regt, Henk. 2017. *Understanding Scientific Understanding.* Oxford: Oxford University Press.

Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67 (4): 559–79.

Dupré, John. 1984. "Probabilistic Causality Emancipated." *Midwest Studies in Philosophy* 9: 169–75.

———. 2013. "Living Causes." *Aristotelian Society Supplementary Volume* 87 (1): 19–37.

Elgin, Catherine. 2007. *True Enough.* Cambridge, Massachusetts: MIT Press.

Fehr, Carla. 2004. "Feminism and Science: Mechanism Without Reductionism." *NWSA Journal* 16 (1): 136–56.

Gehlert, Sarah, Dana Sohmer, Tina Sacks, Charles Mininger, Martha McClintock, and Olufunmilayo Olopade. 2008. "Targeting Health Disparities: A Model Linking Upstream Determinants To Downstream Interventions." *Health Affairs* 27 (2): 339–49.

Hiatt, Robert A. 2004. "The Social Determinants of Cancer." *European Journal of Epidemiology* 19 (9): 821–22.

Howick, Jeremy, Paul Glasziou, and Jeffrey K. Aronson. 2013. "Problems with Using Mechanisms to Solve the Problem of Extrapolation." *Theoretical Medicine and Bioethics* 34 (4): 275–91.

Illari, Phyllis McKay. 2011. "Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis." *International Studies in the Philosophy of Science* 25 (2): 139–57.

Ioannidis, J. P. A., S. D. Schully, T. K. Lam, and M. J. Khoury. 2013. "Knowledge Integration in Cancer: Current Landscape and Future Prospects." *Cancer Epidemiology Biomarkers & Prevention* 22 (1): 3–10.

Krieger, Nancy. 2008. "Proximal, Distal, and the Politics of Causation: What's Level Got to Do With It?" *American Journal of Public Health* 98 (2): 221–30.

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

Mayo-Wilson, Conor. 2014. "The Limits of Piecemeal Causal Inference." *The British Journal for the Philosophy of Science* 65 (2): 213–49.

Mitchell, Sandra D. 2002. "Integrative Pluralism." *Biology & Philosophy* 17: 55–70.

O'Malley, Maureen A, and Karola Stotz. 2011. "Intervention, Integration and Translation in Obesity Research: Genetic, Developmental and Metaorganismal Approaches." *Philosophy, Ethics, and Humanities in Medicine* 6 (1): 2.

Pickersgill, Martyn. 2016. "Epistemic Modesty, Ostentatiousness and the Uncertainties of Epigenetics: On the Knowledge Machinery of (Social) Science." *The Sociological Review* 64 (supplement): 186–202.

Plutynski, Anya. 2018. *Explaining Cancer.* Oxford: Oxford University Press.

Reiss, Julian. 2019. "Against External Validity." *Synthese* 196: 3103–3121.

Russo, Federica, and Jon Williamson. 2007. "Interpreting Causality in the Health Sciences." *International Studies in the Philosophy of Science* 21 (2): 157–70.

Sampson, R. J., Stephen W. Raudenbush, and Felton Earls. 1997. "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy." *Science* 277 (5328): 918–24.

Shavers, Vickie L. 2007. "Measurement of Socioeconomic Status in Health Disparities Research." *Journal of the National Medical Association* 99 (9): 1013–23.

Steel, Daniel. 2016. "Climate Change and Second-Order Uncertainty: Defending a Generalized, Normative, and Structural Argument from Inductive Risk." *Perspectives on Science* 24 (6): 696-721.

———. 2018. "Wishful Thinking and Values in Science." *Philosophy of Science.* 85 (5): 895-905.

Trout, J. D. 2002. "Scientific Explanation and the Sense of Understanding." *Philosophy of Science* 69(2): 212-33.

Woodward, James. 2000. "Explanation and Invariance in the Special Sciences." *The British Journal for the Philosophy of Science* 51 (2): 197–254.

———. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology & Philosophy* 25: 287–318.

To be presented at PSA2020, the 27th Biennial Meeting of the Philosophy of Science Association

Causal Pluralism in Philosophy:

Empirical Challenges and Alternative Proposals

Phuong (Phoebe) Dinh & David Danks

Carnegie Mellon University

Finalized draft as of March 6, 2020

Abstract

An increasing number of arguments for causal pluralism invoke empirical psychological data. Different aspects of causal cognition—specifically, causal perception and causal inference—are thought to involve distinct cognitive processes and representations, and they thereby distinctively support transference and dependency theories of causation, respectively. We argue that this dualistic picture of causal concepts arises from methodological differences, rather than from an actual plurality of concepts. Hence, philosophical causal pluralism is not particularly supported by the empirical data. Serious engagement with cognitive science reveals that the connection between psychological concepts of causation and philosophical notions is substantially more complicated than is traditionally presumed.

1. Introduction

Imagine a billiard ball rolling into a stationary ball, immediately followed by movement of the latter ball. Now imagine a healthy person who brushes against a plant and develops a rash an hour later. While both seemingly causal, these two sequences differ along many dimensions (e.g., timeframe, domain, reliability). In response, cognitive science research arguably points towards at least two distinct concepts of causation, one driven chiefly by perceptual features, the other statistical (e.g., Lombrozo 2010). At the same time, many philosophers of causation have argued—explicitly or otherwise—that the metaphysics of causation should depend partly on its psychological plausibility (e.g., Woodward 2011a, 2011b; Hitchcock 2012). These arguments are not simplistic inferences from psychological to metaphysical reality, but rather an observation that, for example, our causal concepts should be defeasibly anchored in actual relations in the world. The result of these two lines of work is a pluralistic metaphysical picture in which causation not only appears differently, but also comes in "basic and fundamentally different varieties" (Hall 2004, 1; see also Hitchcock 2003).

At a high level, the general argument-schema that unifies many different proposals of causal pluralism can be understood as:

(1) Our lay concept(s) of causation are defeasibly correlated or connected with the metaphysical or scientific relation(s) of causation in the world

(2) Cognitive science tells us that we have multiple distinct lay concepts of causation, realized through distinct cognitive processes and representations

⇒ (Conclusion) Metaphysical or scientific causal pluralism is defeasibly correct

One way to resist this argument is to challenge premise (1) by arguing that our lay concepts need not have any connection with metaphysical or scientific relations. In this paper, we instead

challenge premise (2): we argue that the appearance of multiple causal concepts in human cognition can be explained by methodological variations between communities of cognitive scientists. Moreover, we show that there are empirical data in support of complex interactions between the perceptually- and statistically-driven concepts of causation, thereby suggesting a single (perhaps complex) lay concept of causation. We conclude by showing how different empirically possible theories of causal cognition have different metaphysical implications, and so the need for philosophically-motivated cognitive science to resolve these issues.

2. Causal Pluralism in Cognitive Science

One cannot help but see a flying baseball break the window (not just be correlated with the breakage), or a person running at top speed because of (and not just in conjunction with) a barking dog. The standard cognitive science account of these phenomena (Michotte 1963) is that such impressions of causation result from a perceptually driven concept characterized by signature spatiotemporal features (e.g., spatiotemporal contiguity between purported cause and effect). The resulting causal perception exhibits a set of distinctive features: automatic, phenomenologically instantaneous, unamenable to top-down influences, and highly sensitive to spatiotemporal features. Causal perception has largely been studied through variations on the direct launching paradigm (Michotte 1963; Scholl and Tremoulet 2000): a stationary object *A* is on screen; a moving object *B* enters the screen and moves until it contacts object *A*; at that point, object *B* stops while object *A* moves until it disappears from screen. In ordinary circumstances, participants invariably claim that the moving object kicked, pushed, or launched the stationary object. Notably, a delay between contact and motion, or a gap between the two object at movement onset, destroys any impression of causality. Different spatiotemporal features can

signal different causal processes, though in all cases, causal perception emerges automatically without explicit reasoning.

In contrast, so-called causal inference[1] involves learning about causal relations from information about covariation, contingency, and other statistical information. This information might come from observed correlations (Rottman and Keil 2012) or interventions (Steyvers et al. 2003). For example, one often needs some data or trial-and-error to infer that an infant is crying because of a rash rather than hunger. Direct spatiotemporal connection is not a useful guide for this type of causal cognition (though see below). Strength judgments of a causal relation are sensitive to the degree of covariation between a purported cause and its effect (Shanks and Dickinson 1987; Buehner, Cheng, and Clifford 2003). Statistics also support causal structure learning or the ability to determine how different causal variables relate to one another (Griffiths and Tenenbaum 2005; Lu et al. 2008). In contrast with causal perception, causal inference is: cognitively effortful, has non-salient phenomenology, is largely independent of spatiotemporal features, and strongly amenable to top-down cognitive influences (Buehner and May 2002).

The different behavioral manifestations of the perceptual and statistical concepts of causation—causal perception and causal inference, respectively—are often taken to suggest that these concepts rely on distinct cognitive processes or systems. This suggestion is further supported by a behavioral study (Schlottmann and Shanks 1992), in which participants anecdotally reported that they "knew the collision was not necessary for Object B to move, but

[1] The name is somewhat unfortunate, as causal perception arguably also involves some inferences. Nonetheless, 'causal inference' is the term used in cognitive science to refer to this kind of statistics-driving causal learning.

that it just looked as if it should be" (338). Further evidence for causal pluralism comes from neuroscientific research, which demonstrates a clear differentiation in the brain networks activated during causal perception and causal inference. The perception of causal launching events, compared to that of non-causal launching events, was accompanied by a higher activation level in bilateral V5/MT/MST, the superior temporal sulcus and the left intraparietal sulcus (Blakemore et al. 2001). These areas are involved in complex visual processing, which suggests that causal perception might involve the recovery of causal structures in an event from motion cues (Fugelsang and Dunbar 2009). In contrast, inferential or statistical tasks with causation involved the activation of prefrontal and occipital cortices, precentral gyrus, and parahippocampal gyrus when the data conformed to participants' expectations. A slightly different network—the anterior cingulate, left dorsolateral prefrontal cortex, and the precuneus—was activated when the data were incongruent with expectations (Fugelsang and Dunbar, 2005). Notably, all of these brain areas are typically associated with 'higher' cognition, such as decision-making, conflict resolution, and information integration. Additionally, patients who had a corpus callosotomy (severing the connection between brain hemispheres, usually to treat epilepsy) exhibited a double dissociation between causal perception (seemingly) localizing in the right hemisphere and causal inference (seemingly) in the left hemisphere (Roser et al. 2005). Insofar as one commits to the thesis that different brain network activations imply different brain mechanisms, these neuroscientific results all seem to imply that causal perception and causal inference recruit two different learning mechanisms.

On top of all of these results, causal perception and causal inference also seem to develop at different points during childhood. Humans develop sensitivity for rudimentary cues to causality such as spatial contiguity between 4 and 5½ months of age (Cohen and Amsel 1998), and

perceive direct launching as a causal event based on the appropriate causal roles between 6½ and 10 months of age (Leslie and Keeble 1987; Oakes and Cohen 1990). By 15 months, infants can perceive a three-object causal chain (in which the first object launches the second, which in turn activates the third) as involving a causal relationship between the first and third object (Cohen et al. 1999). In contrast, humans do not successfully solve the blicket detector task—a classic paradigm in causal inference research—until roughly two years of age (Gopnik et al. 2004; Sobel and Kirkham 2006). Children develop more complex causal reasoning abilities, such as the ability to infer unobserved causes (Schulz and Sommerville 2006) and integrate base rates (Griffiths et al. 2011), by four years of age, significantly later than all of the causal perception capacities. The different developmental timelines lend further empirical support to the initial claim that causal perception and causal inference result from two different cognitive processes and representations, which in turn depend on distinct psychological concepts of causation.

3. Causal Pluralism in Philosophy

Many metaphysical accounts of causation can be organized into two clusters of theories: transference and dependency. Transference theorists typically define causation by a transfer of energy or a conservation of quantities through transformation (e.g., Salmon 1984; Dowe 1992, 2000), all of which have signature spatiotemporal properties. For example, in a collision event between two billiard balls, spatiotemporal contiguity during contact enables a transfer of momentum from each ball to the other. In contrast, dependency accounts of causation characterize (though not necessarily define) a causal relation between two factors by their statistical relationship: generative causes make their effect more likely; preventative causes make their effect less likely. This statistical dependency is then grounded in different ways by different

authors, such as counterfactuals, hypothetical interventions, or statistical differences in appropriate reference classes (e.g., Lewis 1973; Woodward 2009).

Rather than arguing for one type of metaphysical account to the exclusion of the other, causal pluralists argue for the co-existence of transference and dependency as distinct kinds of causation that govern different phenomena (Godfrey-Smith 2009). Although arguments for these accounts can be purely metaphysical, a significant subset derive force from the empirical plausibility of these accounts (e.g., Hitchcock 2003; Woodward 2006, 2011a, 2011b; Hall 2004; Lombrozo 2010). The empirical observation that humans exhibit different behaviors and shift their criteria for causation in different scenarios, as presented above, is ostensibly a natural consequence of distinct philosophical causal concepts. For example, there is a straightforward mapping between the signature criteria for causation in many transference accounts and the specific spatiotemporal conditions encoded in the perceptually realized concept of causation. In particular, the immediacy and experiential richness of causal perception is inexplicable by many existing dependency theories of causation, but is more readily explained if causation involves transference (Wolff 2008; Beebee 2009). In the other direction, causal inference can often take place even in the absence of any obvious transfer of force or quantity between agent and recipient, such as cases of prevention in which there is no direct physical connection at all. This kind of causal cognition is often taken to support a dependency notion of causation (Woodward 2009). Causal pluralism seems to explain a wide variety of human causal intuitions—those of both laypeople and philosophers—at the cost of only a slightly more crowded ontology.

4. Causal Pluralism in Cognitive Science: A Methodological Analysis

8

The previous sections presented the "standard view" in cognitive science of causal learning as consisting of two distinct cognitive processes and representations, and noted the important role that it plays in philosophical arguments about the nature of causation. In this section, we challenge the premise that the standard (cognitive science) view is well-supported. In particular, we contend that there are three key methodological differences between causal perception and causal inference experiments, and those differences can explain the differences in observed phenomena without appeal to multiple cognitive processes.

First, the presentation formats and response measures drastically differ between causal perception and causal inference experiments. In the former, participants usually watch a single event and answer questions about that particular event (Scholl and Nakayama 2002). Additionally, participants typically answer a forced-choice question of whether a causal relation exists, or give a quantitative rating of the extent to which a purported causal relation exists in this particular event. In contrast, causal inference experiments involve trial-by-trial presentations of cases (Fernbach and Sloman 2009) or a contingency table summarizing those cases (Hagmayer and Waldmann 2002). Moreover, the typical causal inference measures include (but are not limited to): ratings of proportions in sets of counterfactuals; numeric ratings of the strength of the cause; a categorical choices between causal models; construction or drawing of causal graphs; measures of intervention choices or post-intervention predictions; and more. Neuroscientific and developmental research on causal perception and causal inference typically use the same kinds of stimuli and measures as the corresponding cognitive/behavioral studies. Hence, the stimuli and measures lead directly to phenomenological and behavioral differences without any strong empirical justification, and—as we show next—vastly divergent theories.

Second, our theories of causal perception and causal inference aim to explain judgments about different things, but without directly considering whether they are distinct processes. Causal inference theories aim to explain participant judgments that are based on multiple cases, and that generalize to future instances. In contrast, causal perception theories aim to explain judgments about single cases with no expectation of generalization. That is, causal perception judgments are about token events whereas causal inference judgments are about types. The very construction of the theories thus precludes direct comparison, since they do not attempt to explain the same phenomena. Moreover, the focus of each theory closely correlates with its experimental methods: experientially rich, automatic causal perception can only be captured by judgments of singular events; explicit statistical causal inference can only occur in judgments of multiple instances. The appearance of empirical difference between causal perception and causal inference can be explained by these methodological and focus divergences without any need to appeal to underlying differences in concepts or representations.

A third set of issues arises for the neuroscientific studies, which one might have thought to be immune from the other two worries. Those studies all relied on subtraction methods in which the brain activation map of a null condition is subtracted from that of the experimental conditions. This calculation reveals areas that are uniquely activated in the experimental conditions, yet omits areas that are commonly activated across conditions. Additionally, that an area activates more strongly during causal inference does not mean it is unactivated during causal perception, and vice versa. Cognitive scientists need a thorough investigation of the common areas of activation and the interactions between different brain regions during causal perception and causal inference. Liberal interpretations of early neuroimaging data without sensitivity to these

nuances could produce an overstated conclusion about distinct underlying brain networks for causal perception and causal inference.

5. Cognitive Science in Support of Causal Monism

A careful look at the cognitive science not only undermines the putative inference for pluralism in causal cognition, but actually provides some positive evidence for monism in causal cognition. In particular, consider causal inference and reasoning studies that use both mechanical and statistical information. For example, Kushnir and Gopnik (2007) tested 4-year-old children with a modified blicket detector task in which (i) statistical information matched standard blicket studies in which children infer causality; but (ii) objects were held over the machine rather than placed on it. Children in that task were more inclined to say that such objects were *not* related to the machine's activation; they smoothly integrated spatiotemporal information and constraints into a (seeming) causal inference task. Similarly, Schlottmann (1999) introduced 5-, 7-, 9-, and 10-year-old children to two systems whose inner mechanisms were hidden from view, but described as different. The experimenter dropped one ball (A) into one end of the system, followed by another ball (B) after 3 seconds; the bell rang roughly 1 second after that (i.e., the observed sequence was A-pause-B-bell). The experimenter then showed children the system mechanisms: the fast system had a two-arm seesaw system that rings the bell almost immediately; the slow system had a downward ramp along which the ball had to roll. Intuitively, the children should make different inferences about which ball caused the bell ring: ball A in the slow system and ball B in the fast system. Most children could diagnose the likely mechanism when only one ball was dropped, but 5- and 7-year-old children had difficulty predicting a delay in ringing even when they saw that the slow mechanism was at work. That is, children's causal

learning and inference up to 7 years of age depended on both statistical data accumulated over trials and the spatiotemporal contiguity cues of the events.

Spatiotemporal contiguity shapes causal judgments even among adults. In a series of experiments, Buehner and May (2002) tested the effect of prior knowledge about the time course of a causal relation on their causal judgments of observed contingencies. Participants were given two scenarios: either a light switch immediately causes a light bulb to turn on, or a grenade launch leads to detonation only after a delay. Experiment 1 used a within-subjects design, meaning that participants completed all experimental conditions. Prior experimentation suggested that these two scenarios produced different explicit assumptions about the causal time course, but results showed almost no difference between participants' causal ratings of the light switch and of the grenade launcher: as the delay period between purported cause and effect increased, participants' causal ratings of the cause decreased regardless of the domain or cover story. Buehner and May's preferred explanation is that the 'pull' of the temporal contiguity in the light switch scenario was so strong that it skewed participants' ratings and overshadowed their assumptions about delayed timeframes in the grenade condition. Even if their explanation is incorrect, spatiotemporal cues clearly play a significant role in adult causal inference.

If the previous section provides methodological reasons for doubting the empirical distinction between the perceptual and statistic concepts of causation, this section offers evidence that these concepts are more interconnected than previously thought. We suggest that a pluralism of psychological causal concepts is currently unwarranted by the data, and cautiously propose that monism should again be a feasible theoretical candidate. This monism clearly must allow different types of input, ranging from spatiotemporal cues to frequency and contingency information, but those could lead to distinct behaviors in light of variation in information and

task demands. Admittedly, many details remain to be provided about this kind of monism, but premise (2) in the original argument-schema clearly does not hold as straightforwardly as has been assumed in the philosophical literature. We now turn to the philosophical import of different possible ways of developing a monist account of causal cognition and learning.

6. Philosophical Implications of Alternative Cognitive Accounts

If we do not accept the standard view of pluralism in causal cognition, then we should consider some plausible alternatives. First, we could prioritize the developmental data, which suggest that the perceptual notion of causation develops first, and then the statistical or dependency notion emerges from it. As young learners gain perceptual exposure to simple causal events such as collision or pulling, their mental representations of these events include relevant perceptual features. Further exposure to new instances can result in automaticity of processing, which thereby manifests behaviorally as causal perception. For more complex causal events, some of the causal representations might not bear the same perceptual characteristics as those previously learned. Patterns of characteristics over multiple token perceptions can provide input to the later-developing statistical notion of causation, which are originally grounded in abstraction over spatiotemporal features. The resulting two concepts might develop to be distinct in adults; the developmental story underdetermines the final number of causal concepts in adults. On this theory, a transference notion of causation provides the historical basis for all causal judgments, and perhaps the actual conceptual basis if the concepts are *not* independent in adults. That is, a true causal relation must involve either a transference of energy or a preservation of some quantity from state to state during transformation, where those spatiotemporal features might be imputed on a system from statistical data.

Second, the order of development of causal cognition could be reversed: humans might develop the statistical notion of causation before the perceptual one. In particular, if causal inference occurs implicitly, then repeated exposures to a causal relation could enable learners to construct a representation of this causal relation that encodes statistical information, such as the frequency of occurrence of the purported effect (base rate), the strength of covariation between cause and effect, and so on. In practice, many of these relations are highly reliable or completely deterministic, and all exhibit reliable spatiotemporal characteristics (e.g., contiguity). Thus, the perceptual features of spatiotemporal contiguity might be encoded alongside—and perhaps even stand in for—statistical information: seeing that a rolling ball makes contact with a stationary ball is sufficient for the prediction that the stationary ball is likely to start moving. A philosophical monist account inspired by this psychological picture would take the dependency notion of causation as fundamental. Notably, causation in the physical, mechanical world typically has (statistically) reliable features such the appearance of determinism, or the ubiquity of spatiotemporal contiguity. The dependency framework can thus account for cases that are typically characterized by spatiotemporal features alone (such as causal perception): those features indicate an underlying, deterministic causal relation (see also Woodward 2011a).

Third, a unitary concept of causation might underpin all of human causal cognition. Importantly, this underlying concept is irreducible to either the perceptual or the statistical concept alone, but is rather *inferred* from features of the seemingly distinct types of cognition. In this theory, people infer the existence of an unobserved causal relation using essentially anything that might be relevant, whether spatiotemporal constraints, information about mechanisms, or reliability of control interventions. One can even imagine other types of information being relevant in one's causal judgment, such as color: the color of a mushroom might suggest its

toxicity. These different information sources are integrated to infer the causal connection (if any), which could potentially require significant tradeoffs by the cognizer. Critically, in the case of events where people have had substantial 'perceptual practice' such as collision events, the spatiotemporal contiguity of the event might be the most important. In the case of events with only statistical information (e.g., determining if the peanuts or the shrimps caused an allergic reaction), people can use other types of information to both infer the relation and to suggest alternative control actions. On this account, causal perception and causal inference are only different behavioral manifestations—in response to different task demands and stimuli types—of the same process(es) and mechanisms whose goal is to yield usable representations of the causal web of the world. To the extent that a unitary concept of causation is psychologically plausible and distinct from the previous two alternatives, the third resulting monistic account of causation might plausibly depart from transference and dependency accounts in various ways. Instead of solely relying on transference or dependency between cause and effect, causation can be characterized by both, and which features are more salient depend on how epistemically accessible they are. In the case of billiard balls colliding into each other, the epistemically accessible features are the contact between the balls and the immediacy with which they move differently upon contact. In the case of prevention (e.g., plugging a hole in the sink prevents the leakage), the epistemically accessible features include the absence of leakage after plugging, and that leakage continues when one fails to plug the hole.

7. Conclusion

Many variants of causal pluralism in philosophy, most of which lean on the distinction between transference-based and dependency-based causation, map onto the parallel development

of two research clusters in cognitive science: the perceptually- and statistically-driven concepts of causation, respectively. In this paper, we argued that the foundation of the dualistic research program in cognitive science is shaky insofar as it traces an artifactual divergence between research paradigms and measures, rather than a natural fault line in the empirical landscape of causal cognition. We then discussed how cognitive science might point to a version of causal monism that includes interactions between the perceptual and statistical concepts of causation. Finally, we sketched three alternative accounts of how the statistical and perceptual concepts of causation might relate to each other, and briefly discussed the implications each of those account would have on the philosophical picture of causation. This sketch is not an endorsement of any particular theory, and the three alternatives are non-exhaustive: There are certainly other possibilities that we cannot explore here due to space limits. Rather, the sketch is an invitation for philosophers and psychologists alike to consider these un- and under-explored alternatives before settling for any particular theory.

References

Beebee, Helen, Christopher Hitchcock, and Peter Menzies, eds. 2009. *The Oxford Handbook of Causation*. Oxford University Press.

Beebee, Helen. 2009. "Causation and Observation." In Beebee et al. 2009, 471–97.

Blakemore, Sarah-Jayne, Pierre Fonlupt, Mathilde Pachot-Clouard, Céline Darmon, Pascal Boyer, Andrew N. Meltzoff, Christoph Segebarth, and Jean Decety. 2001. "How the brain perceives causality: an event-related fMRI study." *Neuroreport* 12(17):3741–46.

Buehner, Marc J., Patricia W. Cheng, and Deborah Clifford. 2003. "From covariation to causation: a test of the assumption of causal power." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(6):1119–40.

Buehner, Marc J., and Jon May. 2002. "Knowledge mediates the timeframe of covariation assessment in human causal induction." *Thinking and Reasoning* 8(4):269–95.

Cohen, Leslie B., and Geoffrey Amsel. 1998. "Precursors to infants' perception of the causality of a simple event." *Infant Behavior and Development* 21(4):713–31.

Cohen, Leslie B., Leslie J. Rundell, Barbara A. Spellman, and Cara H. Cashon. 1999. "Infants' perception of causal chains." *Psychological Science* 10(5):412–18.

Dowe, Phil. 1992. “Process causality and asymmetry.” *Erkenntnis* 37(2):179–96.

———. 2000. *Physical Causation*. Cambridge: Cambridge University Press.

Fernbach, Philip M., and Steven A. Sloman. 2009. "Causal learning with local computations." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35(3):678–93.

Fugelsang, Jonathan A., and Kevin N. Dunbar. 2005. "Brain-based mechanisms underlying complex causal thinking." *Neuropsychologia* 43(8):1204–13.

———. 2009. “Brain-based mechanisms underlying causal reasoning.” In *Neural correlates of thinking. On thinking,* vol. 1, ed. Eduard Kraft., Balázs Gulyás, and Ernst Pöppel, 269–79. Springer, Berlin, Heidelberg.

Godfrey-Smith, Peter. 2009. “Causal pluralism.” In Beebee et al. 2009, 326–37.

Gopnik, Alison, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks. 2004. "A theory of causal learning in children: causal maps and Bayes nets." *Psychological Review* 111(1):3–32.

Griffiths, Thomas L., David M. Sobel, Joshua B. Tenenbaum, and Alison Gopnik. 2011. "Bayes and blickets: Effects of knowledge on causal induction in children and adults." *Cognitive Science* 35(8):1407–55.

Griffiths, Thomas L., and Joshua B. Tenenbaum. 2005. "Structure and strength in causal induction." *Cognitive Psychology* 51(4):334–84.

Hagmayer, York, and Michael R. Waldmann. 2002. "How temporal assumptions influence causal judgments." *Memory and Cognition* 30(7):1128–37.

Hall, Ned. 2004. "Two concepts of causation." In *Causation and Counterfactuals*, ed. John David Collins, Edward J. Hall, and Laurie Ann Paul, 225–76. Cambridge, Massachusetts: MIT Press.

Hitchcock, Christopher. 2003. “Of Humean Bondage.” *British Journal for the Philosophy of Science* 54:1–25.

———. 2012. "Portable causal dependence: A tale of consilience." *Philosophy of Science* 79(5):942–51.

Kushnir, Tamar, and Alison Gopnik. 2007. “Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions.” *Developmental Psychology* 43:186–96.

Leslie, Alan M. and Stephanie Keeble. 1987. Do six-month-old infants perceive causality?. *Cognition* 25(3):265–88.

Lewis, David. 1973. “Causation.” *Journal of Philosophy* 70(17):556–67.

Lombrozo, Tania. 2010. “Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions.” *Cognitive Psychology* 61(4):303–32.

Lu, Hongjing, Alan L. Yuille, Mimi Liljeholm, Patricia W. Cheng, and Keith J. Holyoak. 2008. "Bayesian generic priors for causal learning." *Psychological Review* 115(4):955–84.

Michotte, Albert. 1963. *The Perception of Causality*. Trans. T. R. Miles and E. Miles. Andover, Hants: Methuen. Original work published in 1946.

Oakes, Lisa M., and Leslie B. Cohen. 1990. "Infant perception of a causal event." *Cognitive Development* 5(2):193–207.

Roser, Matthew E., Jonathan A. Fugelsang, Kevin N. Dunbar, Paul M. Corballis, and Michael S. Gazzaniga. 2005. Dissociating processes supporting causal perception and causal inference in the brain. *Neuropsychology* 19(5):591–602.

Rottman, Benjamin M., and Frank C. Keil. 2012. "Causal structure learning over time: Observations and interventions." *Cognitive Psychology* 64(1-2):93–125.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, New Jersey: Princeton University Press.

Schlottmann, Anne. 1999. "Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism." *Developmental Psychology* 35(1): 303–17.

Schlottmann, Anne, and David R. Shanks. 1992. "Evidence for a distinction between judged and perceived causality." *The Quarterly Journal of Experimental Psychology* 44(2):321–42.

Scholl, Brian J., and Patrice D. Tremoulet. 2000. "Perceptual causality and animacy." *Trends in Cognitive Sciences* 4(8):299–309.

Scholl, Brian J., and Ken Nakayama. 2002. "Causal capture: Contextual effects on the perception of collision events." *Psychological Science* 13(6):493–98.

Schulz, Laura E. and Jessica Sommerville. 2006. God does not play dice: Causal determinism and children's inferences about unobserved causes. *Child Development* 77(2):427–42.

Shanks, David R. and Anthony Dickinson. 1987. "Associative accounts of causality judgment." *The Psychology of Learning and Motivation* 21:229–61.

Sobel, David M., and Natasha Z. Kirkham. 2006. "Blickets and babies: the development of causal reasoning in toddlers and infants." *Developmental Psychology* 42(6):1103–15.

Steyvers, Mark, Joshua B. Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. 2003. "Inferring causal networks from observations and interventions." *Cognitive science* 27(3):453–89.

Wolff, Phillip. 2008. "Dynamics and the perception of causal events." In *Understanding Events: From Perception to Action*, ed. Thomas F. Shipley and Jeffrey M. Zacks, 555–86. Oxford University Press.

Woodward, James. 2006. "Sensitive and insensitive causation." *The Philosophical Review* 115(1):1–50.

———. 2009. “Agency and Interventionist Theories.” In Beebee et al. 2009, 234–62.

———. 2011a. "Causal perception and causal cognition." In *Perception, Causation, and Objectivity*, ed. Johannes Roessler, Hemdat Lerman, and Naomi Eilan, 229–63. Oxford University Press.

———. 2011b. "Psychological studies of causal and counterfactual reasoning." *Understanding Counterfactuals, Understanding Causation. Issues in Philosophy and Psychology*, ed. Christoph Hoerl, Teresa McCormack, and Sarah R. Beck, 16–53. Oxford University Press.

The substantial role of Weyl symmetry in deriving general relativity from string theory

John Dougherty

November 8, 2020

Abstract

String theory reduces to general relativity in appropriate regimes. Huggett and Vistarini have given an account of this reduction that includes a deflationary thesis about symmetry: though the usual derivation of general relativity from string theory appeals to a premise about the theory's symmetry, Huggett and Vistarini argue that this premise plays no logical role. In this paper I disagree with their deflationary thesis and argue that their analysis is based on a popular but flawed conception of the interaction between symmetry and quantization. On this conception, quantization can break symmetries of the classical theory, and we must decide whether these symmetries should be reimposed. I argue that a better conception recognizes a tripartite distinction between ordinary, broken, and anomalous symmetries.

1 Introduction

The basic conceptual task for a quantum theory of gravitation is to recover something like the rough-and-ready picture of space and time that we use to characterize the target gravitational phenomena. Much recent philosophical work on this topic addresses general questions about this task: the extent to which theories must presuppose the rough-and-ready picture, the different possible success conditions for a recovery, whether and how to make sense of claims that spacetime emerges from some more fundamental quantum features of the world, and so on (Huggett and Wüthrich, 2013; Crowther, 2018). But there are also specific questions that arise within particular research programs. For example, Huggett and Vistarini (2015) and Vistarini (2019) point out that symmetry considerations seem to be "a key concept connecting string theory to phenomenological space-time" (2015, 1170) but that the status of these considerations is obscure. Despite the apparent importance of symmetry, Huggett and Vistarini argue that it is merely a formal feature of string theory, suggesting that it can play no substantial role. Getting to philosophical grips with string theory as a quantum theory of gravity requires a resolution of this conceptual tension.

This paper disagrees with Huggett and Vistarini's account and suggests an alternative. Their argument focuses on the interaction between symmetry and quantization. We recover the Einstein field equation (EFE) in string theory by quantizing a classical theory that exhibits so-called Weyl symmetry. On Huggett and Vistarini's telling, quantization breaks this Weyl symmetry, and we face a choice: if we decide to reimpose the symmetry then we are led to the EFE, and if we try to leave the symmetry broken then it will reappear in a different guise and again lead to the EFE. They conclude that Weyl symmetry is unavoidable and hence "not a logically independent postulate" (2015, 1173) of string theory. But I think this framing is misleading. When a theory is quantized, its symmetries have three possible fates: they might be preserved, they might be broken, or they might be anomalous. As I will argue, Huggett and Vistarini confine their attention to the first case; that is, their conclusion that Weyl symmetry is always preserved rests on considering only those cases in which it is preserved. Further attention to the other cases—and especially to cases in which the symmetry is anomalous—shows that Weyl symmetry is not a merely formal feature of string theory. It also illustrates the more general use of symmetries in the string-theoretic approach to theory construction, which plays an important role in defenses of the string theory program (Dawid, 2013).

The plan is as follows. In Section 2 I isolate the feature of Huggett and Vistarini's framing that I disagree with. Though this debate is motivated by the string-theoretic derivation of the EFE, my disagreement with Huggett and Vistarini is really a disagreement about the interpretation of quantum field theories with spacetime-dependent symmetries. It bears on their analysis only insofar as the derivation they discuss occurs within such a theory. As I argue, their presentation supposes that classical symmetries are either broken or preserved in the process of quantization. Section 3 argues that this is not a natural dichotomy, for there is a third possibility: the symmetry might be anomalous. Anomalous symmetries are in some sense preserved and in some sense broken; as such, they fit uncomfortably in Huggett and Vistarini's account. Section 4 uses this third category to argue that Weyl symmetry is not merely a formal feature of string theory.

2 Weyl symmetry

The argument that string theory reproduces general relativity in the appropriate domains has various prongs; Weyl symmetry is primarily relevant to the recovery of the EFE for the spacetime metric. In the appropriate regimes, a morass of string excitations ought to look like a Lorentzian metric to a test string moving through it. If the test string is to have a consistent quantization it must be Weyl invariant, and if it is to be Weyl invariant then the effective Lorentzian metric must satisfy the EFE. Or at least, this is the standard story. Huggett and Vistarini ultimately argue that this appeal to Weyl invariance can be circumvented: whether or not we suppose Weyl invariance, the EFE will follow. This is the claim I want to take issue with.

On Huggett and Vistarini's analysis, string theory's recovery of general relativity—and, thereby, phenomenological space—is expressed by two results. First, the spectrum of the string contains gravitons, the force carriers for gravity. More precisely, upon quantizing the string you will find quantum states containing massless spin-2 particles, the representation of the Lorentz group in which gravitons live. Second, an adequate quantization of a string moving in an approximately classical background made up of these massless spin-2 quanta requires the background to satisfy the EFE. String theory therefore contains the right stuff behaving in the right way to reproduce general relativity in the right regimes.

This paper is concerned with the second of these results, which is set within an effective theory of a string propagating in a classical background. Effective field theories model the salient degrees of freedom in systems where the fundamental degrees of freedom are unknown, or cannot be connected to the salient degrees of freedom, or are computationally intractable. For example, the Standard Model of particle physics contains twelve elementary matter particles. While some of these particles are observable in isolation at low energies, some only appear in bound states—the up and down quarks only occur as constituents of protons, neutrons, and pions. The effective degrees of freedom at low energies are therefore not those appearing in the Standard Model, and the effective field theory used to model physics at this scale is formulated directly in terms of protons, neutrons, and pions instead of quarks. Analogously, the EFE is derived from string theory in an effective theory that replaces gravitonic excitations with a classical Lorentzian metric. It's this metric that must satisfy the EFE.

More formally, the effective theory of interest is the following. The classical background is given by a metric G on a manifold X of dimension D. A possible history for a string is a map $\Sigma \to X$ with Σ a two-dimensional surface. We define a quantum field theory on Σ with two fluctuating fields: a Lorentzian metric g on Σ and a map $\phi : \Sigma \to X$ picking out a possible history for the string. The action for this theory is

$$S(g, \phi) = \int_\Sigma \|d\phi\|^2 \, \mathrm{vol}_g$$

where vol_g is the volume element on Σ determined by g and the norm is induced by g and G. That is, picking some coordinates on Σ and X, we have

$$\|d\phi\|^2 = g^{mn} \left(\partial_m \phi^\mu\right) \left(\partial_n \phi^\nu\right) G_{\mu\nu} \qquad\qquad \mathrm{vol}_g = \sqrt{-|g|}\, d^2x$$

where Roman indices run over the two dimensions of Σ and Greek indices over the D dimensions of X.

The EFE for G is obtained by requiring the quantum theory to be well-behaved. As a first pass at articulating this requirement, consider the path integral quantization of the action above. The theory is determined by the path integral

$$\int \mathcal{D}g \int \mathcal{D}\phi \, \exp(iS(g, \phi))$$

If we fix a metric g on the worldsheet Σ, the inner integral is the path integral for a quantum field theory of D scalar fields in two spacetime dimensions. Theories of this kind are relatively well understood, and the integral over ϕ is relatively easy to perform, at least when G is nearly flat. The path integral of our effective theory is therefore an integral over a family of scalar field theories indexed by metrics g on Σ. Integrating out the scalar fields, our integral becomes

$$\int \mathcal{D}g \, \exp(iS_{\text{red}}(g))$$

where $S_{\text{red}}(g)$ is an action for g that incorporates the quantum fluctuations of the field ϕ. The full path integral ought therefore reduce to an integral over g alone.

We can only integrate out the field ϕ if G satisfies the EFE. If the reduced action S_{red} exists then its exponentiation must have the same symmetries as the integral over ϕ. In particular, note that the original action $S(g, \phi)$ is invariant under the Weyl transformation

$$g \mapsto e^{2\omega} \, g$$

determined by a positive real-valued function ω on Σ, since the induced change in the norm of $d\phi$ cancels out the induced change in the volume element. On the other hand, to leading order in fluctuations in ϕ, an infinitesimal Weyl transformation with parameter ω shifts S_{red} by (D'Hoker, 1999, Eq. 6.61)

$$\delta S_{\text{red}}(g) = -\frac{1}{2\pi} \int_\Sigma d^2x \, \sqrt{-|g|} \, \omega \left(\frac{1}{2} g^{mn} (\partial_m \phi^\mu)(\partial_n \phi^\nu) R^G_{\mu\nu} + \frac{D-26}{6} R^g \right)$$

with $R^G_{\mu\nu}$ the Ricci tensor associated with the metric G and R^g the scalar curvature associated with the metric g. Since S_{red} must have the same symmetries as the original theory, this shift must vanish for all ω and g. This implies that the first term in the integrand vanishes for all g and ϕ, so that $R^G_{\mu\nu} = 0$. And this is the EFE in vacuum. The argument generalizes: if we add other classical background fields on X to the effective action then the shift in S_{red} under a Weyl transformation will include terms involving these other fields, and the shift will vanish when $R^G_{\mu\nu}$ satisfies the EFE determined by the stress-energy tensor of the added fields.

Huggett and Vistarini argue that Weyl symmetry plays no substantial role in the derivation I've just sketched; this is our point of disagreement. The derivation relied on the claim that S_{red} must have Weyl symmetry, and this "must" requires justification. Huggett and Vistarini argue that it is tautologous:

> although the derivation of the EFEs appeals to [Weyl] symmetry, since that is itself a consequence of string theory, it is not, logically speaking, a necessary premise of the derivation (2015, 1173).

On their view, the appeal to Weyl symmetry could in principle be eliminated. We must have $R^G_{\mu\nu} = 0$ (and $D = 26$), and S_{red} must be Weyl-invariant, but

according to Huggett and Vistarini this is a downstream consequence of other hypotheses in string theory. We could just as well take a different route, one that made no explicit detour through Weyl symmetry.

Weyl symmetry certainly appears to play a role in the derivation just sketched, so Huggett and Vistarini argue for their triviality thesis by arguing that this is a mere appearance. In light of the generally nontrivial behavior of S_{red} under Weyl transformations, Huggett and Vistarini say that the Weyl symmetry of the original action is "broken by quantization" and that the EFE appears to follow when the symmetry is "reimposed" on S_{red} (2015, 1170). If we don't reimpose the symmetry then it seems we might have $R^G_{\mu\nu} \neq 0$ or $D \neq 26$. But, they claim, if we don't demand Weyl invariance then we must change our classical background:

> In this case, different choices of conformal factor in the Weyl transformation of the internal metric... will be physically different. Hence, $[\omega]$ is a new physical degree of freedom over the worldsheet, a scalar background field: specifically a dilaton field $[\Phi]$ (2015, 1171).

Suppose, then, that we adopt a different effective field theory, one that includes a scalar field Φ on X. Huggett and Vistarini argue that if we suppose Φ to be tachyonic, and if we suppose that some mechanism gives it good long-distance behavior, then we can show that Weyl invariance must hold. They conclude that Weyl invariance is unavoidable: even if we suppose that it doesn't hold we can derive that it does.

In the rest of this paper I argue that Huggett and Vistarini's reasoning does not go through and that talk of breaking and reimposing Weyl invariance is misleading. The Weyl symmetry of S_{red} is a necessary premise in the derivation of the EFE just sketched and is not a logical consequence of some other hypotheses. Huggett and Vistarini's argument only shows that S_{red} is Weyl-invariant under the hypothesis that S_{red} is Weyl-invariant. The triviality of this conclusion is obscured by a common way of talking about the role of symmetries in quantization, according to which quantization can break symmetries of the classical theory and it's left for us to decide whether to reimpose them. A better accounting of the situation distinguishes between cases in which the symmetry is preserved, cases in which it cannot be implemented, and cases in which it is anomalous.

3 Anomalies

My disagreement with Huggett and Vistarini's framing isn't particular to Weyl symmetry but applies to symmetries of all kinds. This section illustrates an alternative framing according to which any symmetry might be preserved, broken, or realized anomalously. Huggett and Vistarini also take their discussion to generalize to other kinds of symmetry. They explicitly analogize Weyl and gauge symmetry, and elsewhere Vistarini suggests that the possibility of a substantial role for Weyl symmetry "challenges the general idea that gauge symmetries are simply formal features of the way in which a theory's physical content is

formally represented" (2019, 40).[1] I agree that deflationary views about Weyl and gauge symmetry stand and fall together, but I think they are untenable in both cases. They fail to mark an important distinction between anomalous and broken symmetries.[2]

3.1 Anomalous global symmetries in field theories

There's an important difference between a theory's being invariant under a symmetry or anomalous, and both of these situations are importantly different from a theory lacking that symmetry altogether. These differences can be illustrated by simpler theories with obvious physical application. Anomalous symmetries are also found in quantum field theory, where they can play an important role in saving the phenomena. The chiral anomaly in the Standard Model is a relatively simple example that illustrates why anomalous global symmetries are acceptable.

A particularly simple instance of the chiral anomaly appears in quantum electrodynamics with one charged fermion. The setting is four-dimensional Minkowski space, and the two fields in the theory are the electromagnetic gauge potential A and a massless Dirac fermion ψ. The action is

$$S(A,\psi) = \int_{\mathbb{M}^4} d^4x \left(-\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \bar{\psi} \not{D} \psi \right)$$

with $F_{\mu\nu}$ the field strength and $\not{D}$ the Dirac operator determined by A. As in the string theory of Section 2, the theory is specified by an iterated path integral

$$\int \mathcal{D}A \int \mathcal{D}\psi \, \mathcal{D}\overline{\psi} \, \exp(iS(A,\psi))$$

As before, we think of the inner integral as defining a quantum field theory with a single fluctuating fermion field ψ in the presence of a fixed classical electromagnetic potential A. And again as before, we proceed by integrating out the fermionic degrees of freedom to obtain an action depending only on A, with the full theory given by performing this remaining integration.

The integral over the fermion field transforms anomalously under the global symmetries of the action. Recall that a Dirac fermion ψ naturally decomposes into two parts: a left-handed Weyl fermion and a right-handed Weyl fermion. The Dirac operator $\not{D}$ is chirally symmetric, so the fermion term in $S(A,\psi)$ can be split into two terms, one involving the left-handed component of ψ and one involving the right-handed component. The action $S(A,\psi)$ is therefore invariant under two kinds of phase transformations,

$$\psi \mapsto e^{i\theta} \psi \qquad\qquad \psi \mapsto e^{i\theta\gamma^5} \psi$$

[1] See Healey (2007) and Redhead (2003) for more detailed articulations of this general idea as well as some discussion about how issues of symmetry and quantization are related to more obviously philosophical issues.

[2] What follows are two simple examples of anomalies. See Monnier (2019) for more thorough but still relatively informal discussions of anomalous quantum field theories.

the first of which rotates the phase on the left- and right-handed components by the same angle θ, and the second of which rotates the phases of each component the same magnitude θ but in opposite directions. Call the latter a chiral phase rotation, since it treats left- and right-handed components differently. While the integral over the fermion fields invariant under the first type of phase rotation, a chiral rotation by θ shifts the quantum effective action by

$$\delta S_{\text{eff}}(A, \psi) = -\frac{Q^2}{16\pi^2} \int_{\mathbb{M}^4} d^4x \, \theta \epsilon^{\mu\nu\alpha\beta} F_{\mu\nu} F_{\alpha\beta}$$

where Q is the charge of the fermion. The effective fermion action transforms under chiral rotations, so there's a sense in which it exhibits chiral symmetry. But it is not invariant; it has an anomaly.

The chiral anomaly doesn't vanish, but this isn't a problem. The derivation in Section 2 required the Weyl anomaly to vanish, but a vanishing chiral anomaly would lead to empirical inadequacy (Dougherty, 2020). For example, neutral pions decay to photons at a rate proportional to the chiral anomaly. If the chiral anomaly vanished then pions would hardly ever decay to two photons, but this is their most common decay channel. Indeed, the chiral anomaly was first discovered when trying to account for the neutral pion's decay rate. As another example, the mass of the η' meson is approximately proportional to the chiral anomaly. A theory without the chiral anomaly gets the η' meson's mass wrong by almost an order of magnitude.

The effective action varies under chiral rotations, but it exhibits chiral rotation symmetry in a weaker sense. Because the chiral anomaly is a reflection of this weaker invariance, it reflects the structure of the symmetry group by satisfying the so-called Wess–Zumino consistency conditions. This is importantly different from a theory that is not invariant under chiral rotations at all, like a theory with massive fermions. These two cases should be distinguished.

3.2 Anomalous gauge symmetries in field theories

Anomalous global symmetries like the chiral symmetry of Section 3.1 or Galilei symmetry in nonrelativistic quantum mechanics are unobjectionable. Indeed, they are desirable, because neutral pions often decay and the mass of a non-relativistic particle isn't state-dependent. Anomalous spacetime-dependent symmetries are less anodyne. These include Weyl symmetry when $R^G_{\mu\nu} \neq 0$ or $D \neq 26$, but they are also found in minor modifications of the Standard Model. The demand for a vanishing Weyl anomaly is analogous to the demand for a vanishing gauge anomaly, and the latter is perhaps more easily interpreted in physical terms by comparison with the Standard Model.

To illustrate gauge anomalies, consider a slightly different theory of charged matter. Replace the Dirac fermion in the action of Section 3.1 with a charged left-handed Weyl fermion χ to give the action $S(A, \chi)$. This action exhibits a spacetime-dependent symmetry: for any real-valued function α on $\mathbb{M}^4$ the transformation

$$A_\mu \mapsto A_\mu - \partial_\mu \alpha \qquad\qquad \chi \mapsto e^{iQ\alpha} \chi$$

leaves the action $S(A, \chi)$ unchanged. But when we integrate out the fermion χ the reduced action transforms anomalously:

$$\delta S_{\text{red}}(A) = -\frac{Q^3}{96\pi^2} \int_{\mathbb{M}^4} d^4x \, \alpha \epsilon^{\mu\nu\alpha\beta} F_{\mu\nu} F_{\alpha\beta}$$

This resembles the chiral anomaly but is distinct. The action $S(A, \chi)$ doesn't have chiral symmetry at all—neither ordinary nor anomalous—because it contains only a left-handed Weyl fermion. And integrating out the Dirac fermion ψ from the action $S(A, \psi)$ of Section 3.1 produces a gauge-invariant reduced action, not one that transforms anomalously under gauge transformations.

The Standard Model has no gauge anomalies, and this expresses a nontrivial fact about the charges of its various particles. If we replaced the Weyl fermion in $S(A, \chi)$ with a right-handed fermion of the same charge then we would obtain another anomalous theory, and the anomaly would have the same form but with a sign flip. So we can build a theory with no anomaly if we include two Weyl fermions with the same charge, one of each handedness, for then the two anomaly terms would cancel. This is just the theory of Section 3.1, since a Dirac fermion is a pair of opposite-handed Weyl fermions. By similar reasoning, the anomaly associated with the $U(1)$ hypercharge gauge symmetry in the Standard Model is proportional to (Schwartz, 2014, Eq. 30.73)

$$2(Y_L^3 + 3Y_Q^3) - (Y_e^3 + Y_\nu^3 + 3(Y_u^3 - Y_d^3))$$

where the Ys are the hypercharges of left-handed leptons and quarks and right-handed electron, neutrino, and up- and down-type quarks. In the classical action these charges are freely specifiable independently, but their observed values are such that this expression vanishes. Similar anomaly cancellation conditions hold for other gauge symmetries in the Standard Model. And, of course, these are all cousins of the Weyl anomaly cancellation conditions $R^G_{\mu\nu} = 0$ and $D = 26$.

The examples in this section show that the preserved–broken dichotomy Huggett and Vistarini employ is too coarse a classification. Putting anomalous symmetries in the "broken" bucket neglects the fact that they satisfy nontrivial constraints, like the Wess–Zumino consistency condition. But putting them in the "preserved" bucket erases the difference between cases where anomalies cancel and cases where they don't. In particular, it elides theories where the Weyl anomaly vanishes and theories where it doesn't. Once we recognize that symmetries may be anomalously realized, we can further distinguish between anomalous global symmetries, like chiral symmetry, and anomalous gauge symmetries, like that of an electromagnetically charged Weyl fermion or the Weyl symmetry of Section 2's string theory. While the former obtain in perfectly good theories—both in principle and of particle phenomena—the latter are ruled out in the string-theoretic derivation of the EFE.

4 Theory space

The distinctions introduced in Section 3 clarify the task of justifying the string-theoretic derivation of the EFEs, and they lead to a problem for Huggett and Vistarini's analysis. The desired conclusion of $R^G_{\mu\nu} = 0$ follows from the demand that the Weyl anomaly vanish, and this follows from the demand that the total gauge anomaly always vanish. So we need a justification for this more general demand. I won't try to provide one here. But, supposing this demand is justified, it's a nontrivial one. There are theories that exhibit a nonvanishing gauge anomaly, like electrodynamics with a single Weyl fermion, and there are theories in which the gauge anomaly vanishes. Anomaly cancellation isn't tautologous. Indeed, the stringency of anomaly cancellation is sometimes claimed to uniquely determine a possible model of string theory.

Weyl symmetry plays a substantial role in Section 2's derivation of the EFE. This derivation required the reduced action to be exactly Weyl symmetric, not anomalously so. This is a substantial requirement because it forces us to have $R^G_{\mu\nu} = 0$ and $D = 26$. And this requirement is nontrivial because there are metrics that aren't Ricci flat and there are manifolds with dimension other than 26. In just the same way, demanding gauge anomaly cancellation in the theories of Section 3 or in the Standard Model puts nontrivial constraints on the field content and charges. It rules out a theory containing a single charged Weyl fermion, and it requires the electron's charge to be precisely the opposite of the proton's. Far from being a tautology, the vanishing of the Weyl anomaly is a powerful constraint on the construction of a quantum field theory.

The power of the vanishing anomaly condition requires an equally powerful justification, and I think this deserves further philosophical attention. Certainly we can't count every anomaly as pathological, since the chiral anomaly in Section 3.1 is instrumental in reproducing low-energy collider phenomena. But we can demand that every gauge anomaly vanish, and this demand is often made. It is sometimes said that theories with gauge anomalies aren't "coherent" (Dawid, 2013, 12) or "consistent" (Schwartz, 2014, 627), but this isn't obviously right, at least not in the strict sense. The traditional argument for this conclusion claims that gauge anomalies "destroy the renormalizability, and thus the consistency, of the gauge theory" (Bertlmann, 1996, 245). This seems too quick. Plenty of perfectly respectable theories aren't renormalizable, including the effective field theories used to model low-energy collider physics (Weinberg, 1995, §12.3). On the other hand, these effective field theories have unitary truncations at each order in the momenta, while any finite truncation of a gauge theory spoils unitarity. This is not the place to sort out the exact relationship between gauge anomalies and renormalization, but this relationship should be clarified if we would like to better understand the derivation of the EFE in string theory.

Because the vanishing Weyl anomaly is a nontrivial constraint, Huggett and Vistarini's deflationary argument must misfire. The problem with it is clear if we adapt it to a simpler theory with anomalous gauge symmetry, like the theory of the single charged fermion. Their argument, recall, begins by supposing that the reduced action isn't exactly invariant under the gauge symmetry. It's a

matter of mathematical fact that the reduced action transforms under the gauge symmetry; the only question is whether it's invariant or anomalous. If we suppose it's anomalous then the fermion's charge must be nonzero. The theory then has a gauge anomaly and is not invariant under the gauge symmetry. At this point Huggett and Vistarini introduce new degrees of freedom and show that these cancel the anomaly. The analogous move in our charged fermion theory would be the introduction of further Weyl fermions: one fermion with the opposite handedness and the same charge, or two fermions with the opposite handedness and charge $Q/2$, or one Weyl fermion with the same charge and handedness and two with the same charge and opposite handedness, or something like this. The total gauge anomaly in any of these modified theories vanishes, so they are exactly gauge-invariant. But they're also just different theories. Introducing another fermion doesn't make the theory with one fermion consistent, it gives a theory with two fermions. In the same way, Weyl invariance in a theory containing a background scalar field Φ doesn't lead to Weyl invariance in a theory without a background scalar field.

Huggett and Vistarini's reasoning doesn't show that Weyl invariance is a purely formal requirement, but it can be useful in a different way. Anomaly cancellation can be a guide to theory development, because it can suggest modifications for the sake of anomaly cancellation. If you observe a charged Weyl fermion then there must be at least one more out there, because a theory with only one charged Weyl fermion has a gauge anomaly. Anomaly cancellation therefore constrains our exploration of the possible space of theories. Dawid's (2013) account of non-empirical theory assessment promotes this type of constraint to a general method for evaluating scientific theories. If we have reason to believe that the vast majority of theories have gauge anomalies then the fact that we've found some that lack them—the Standard Model, or the string theory with $R^G_{\mu\nu} = 0$ and $D = 26$—is a good sign that we're on the right track. The antecedent is a big "if", but it does seem difficult to construct theories in which all anomalies cancel.

5 Conclusion

I have argued that Weyl symmetry plays a substantial role in the derivation of the EFE in string theory. More precisely, the EFE follows from the hypothesis that the Weyl anomaly vanishes, and this hypothesis isn't empty. An adequate account of the Weyl anomaly requires a conception of symmetry that goes beyond the preserved–broken dichotomy found in Huggett and Vistarini's analysis and more broadly. I have indicated a replacement. On the alternative framing I have provided, the derivation of the EFE rests on the prohibition of gauge anomalies, and the justification of this prohibition should be further investigated. Leaving these details aside, I think Huggett and Vistarini's deflationary argument doesn't work. It finds Weyl invariance in every theory because it responds to failures of Weyl invariance by changing the theory under consideration. Some theories—indeed, most—are not Weyl invariant.

References

Bertlmann, R. A. (1996). *Anomalies in Quantum Field Theory*. Clarendon Press.

Crowther, K. (2018). Inter-theory relations in quantum gravity: Correspondence, reduction, and emergence. *Studies in History and Philosophy of Modern Physics*, 63:74–85.

Dawid, R. (2013). *String Theory and the Scientific Method*. Cambridge University Press.

D'Hoker, E. (1999). String theory. In *Quantum Fields and Strings: A Course for Mathematicians*, volume 2, pages 807–1012. American Mathematical Society.

Dougherty, J. (2020). Large gauge transformations and the strong CP problem. *Studies in History and Philosophy of Modern Physics*, 69:50–66.

Healey, R. (2007). *Gauging What's Real*. Oxford University Press.

Huggett, N. and Vistarini, T. (2015). Deriving general relativity from string theory. *Philosophy of Science*, 82(5):1163–1174.

Huggett, N. and Wüthrich, C. (2013). Emergent spacetime and empirical (in)coherence. *Studies in History and Philosophy of Modern Physics*, 44(3):276–285.

Monnier, S. (2019). A modern point of view on anomalies. *Fortschritte der Physik*, 67(8-9):1910012.

Redhead, M. (2003). The interpretation of gauge symmetry. In Brading, K. and Castellani, E., editors, *Symmetries in Physics: Philosophical Reflections*, chapter 7, pages 124–139. Cambridge University Press.

Schwartz, M. D. (2014). *Quantum Field Theory and the Standard Model*. Cambridge University Press.

Vistarini, T. (2019). *The Emergence of Spacetime in String Theory*. Routledge.

Weinberg, S. (1995). *The Quantum Theory of Fields*, volume 1. Cambridge University Press.

Towards Mechanism 2.1: A Dynamic Causal Approach

Abstract: I propose a dynamic causal approach to characterizing the notion of a mechanism. Levy and Bechtel, among others, have pointed out several critical limitations of the new mechanical philosophy, and pointed in a new direction to extend this philosophy. Nevertheless, they have not fully fleshed out what that extended philosophy would look like. Based on a closer look at neuroscientific practice, I propose that a mechanism is a dynamic causal system that involves various components interacting, typically nonlinearly, with one another to produce a phenomenon of interest.

1. Introduction

The last three decades have witnessed the rise of the so-called *new mechanical philosophy* (NMP) in philosophy of science. The emergence of this NMP was largely motivated by philosophers' realization that, in contrast with the physical sciences where natural laws play a central role in offering explanation, prediction and understanding, the life sciences are best characterized as a hodgepodge of subdisciplines that focus on discovering and investigating mechanisms. Another motive for the NMP's arising is related to the shift from the focus on scientific theories to on scientific practice.

Advocates of the NMP provide philosophers with a new framework for re-examining many pivotal problems in philosophy of science, e.g., scientific explanation, causation, the autonomy of the special sciences, to name just a few. However, even though the NMP has significantly reshaped the landscape of philosophy of science, there is still a long way to go. Recently, many authors have realized that the framework has serious limitations (Brigandt 2013; Levy and Bechtel 2013; Levy and Bechtel 2016). At the heart of these limitations is the fact that previous work tends to center on qualitative aspects of mechanisms and draws on examples primarily from textbooks in cell and molecular biology, while neglects quantitative/dynamic aspects of mechanisms that are reflected in real scientific practice.

Given these limitations, Levy and Bechtel (2016) call for an extended conception

of mechanisms and mechanistic explanation, the so-called 'mechanism 2.0'.[1] Although Levy and Bechtel, among others,[2] point in the right direction (or so I suppose) and highlight several crucial points regarding what the extended philosophy would look like, they have not yet fully developed their proposal. So, I here, following in their footsteps, take up the mission of developing one version of such an extended philosophy and call it 'mechanism 2.1'. My approach, largely inspired by neuroscientific practice, is capable of capturing both the qualitative and quantitative aspects of mechanisms, and dovetails well with real scientific practice.

The essay unfolds as follows. Section 2 briefly describes the NMP, followed by Section 3 where Levy and Bechtel's proposal for 'mechanism 2.0' is introduced. Section 4 proposes a dynamic causal approach to characterizing mechanisms, and Section 5 discusses what philosophical implications it can deliver.

2. The New Mechanical Philosophy

The NMP represents a bundle of closely connected but slightly different ideas

[1] Notice that Levey and Bechtel (2016)'s interest is in expanding the mechanistic explanation framework rather than the conception of mechanisms. However, I think an extended conception of mechanistic explanation must be built upon an extended conception of mechanisms, since the latter is more fundamental. Yet, their project does inform me of how to develop an extended account of mechanisms.

[2] E.g., Kaplan and Bechtel (2011), and Brigandt (2013).

proposed by a number of philosophers concentrating primarily on practice in the life sciences (Bechtel and Richardson 1993; Machamer et al. 2000; Glennan 2002, 2005; Bechtel and Abrahamsen 2005; Bechtel 2006, 2008; Darden 2006; Craver 2007). These philosophers all agree that we place mechanisms on center stage when examining those traditional philosophical questions (e.g., explanation, causation), even though they have not yet reached a consensus on how to philosophically specify the notion of mechanisms. According to one most commonly cited characterization:

> "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." (Machamer et al. 2000, 3)

In characterizing mechanisms, different authors employ different terminologies which reflect their distinct ontological commitments.[3] Setting aside these ontological disputes, nevertheless, they all seem to agree that a mechanism involves four elements: a phenomenon/behavior, components/parts/entities, interactions/activities/operations,

[3] Machamer et al. (2000) take a dualistic stance towards mechanisms, holding that a mechanism is composed of two ontologically different kinds: entities and activities. Bechtel (2006, 2008) also thinks that a mechanism is composed of two different kinds: component parts and component operations. Glennan (2002), by contrast, takes a monist position, holding that a mechanism is composed of parts that interact to produce a phenomenon of interest.

and spatiotemporal organization/structure. Another element, not clearly shown, is also worth mentioning: multilevel hierarchy.

The multilevel hierarchy is manifested by the fact that the component of a mechanism may constitute a sub-mechanism by itself, and that the mechanism may constitute a component of an even bigger mechanism. This also implies that a mechanism's identification hinges on what target phenomenon/behavior is under question. In other words, there is no mechanism *simpliciter*, but only a mechanism *for* a particular phenomenon/behavior. With respect to components and interactions—in terms of Craver (2007)'s constitutively relevant criterion—only those that contribute to producing a particular phenomenon/behavior of the mechanism count as the components and interactions of the mechanism.

This NMP has significant implications for a number of philosophical issues, e.g., explanation. This philosophy advocates a new account of explanation, i.e., mechanistic explanation. According to this account, explaining a phenomenon/behavior (at least in the life sciences) lies in uncovering a mechanism, i.e., uncovering how the various components interact with one another in a spatiotemporally orchestrated manner to produce the phenomenon of interest. Obviously, there is no role for laws to play, and explanation does not proceed in a manner suggested by the covering-law model of scientific explanation.

No doubt, this philosophy's attractiveness essentially comes down to the fact that it goes in concert with the practice in the life sciences. Yet, as many philosophers have pointed out, although this framework has come very close to practice, it does not

come close enough.

3. Mechanism 2.0: Call for An Extension

Recently, many philosophers have cast doubt on the adequacy of the NMP (Bechtel and Abrahamsen 2010, 2013; Brigandt 2013; Levy and Bechtel 2013, 2016). According to these philosophers, the NMP has the following limitations. First, the NMP treats a mechanism as if it is composed of a *linear* causal sequence. However, scientists have recognized that a mechanism can be a very complex network of interacting components that possesses feedback/feedforward loops, whose interactions are typically non-linear and non-sequential. Second, the NMP routinely concentrates on the structural, organizational, and spatial aspects of a mechanism, ignoring that a mechanism is essentially a *dynamic* system within which the parts are changing over time. Third, these two features, linear and non-dynamic thinking, are always associated with a third feature of that philosophy: *qualitative* thinking. This feature is clearly illustrated by the way the new mechanists qualitatively describe how a mechanism is brought about, and by the simple paradigmatic examples drawn from textbooks (e.g., the *lac* operon of *E. coli*). These qualitative characterizations of mechanisms may help unravel some qualitative aspects of the mechanism, but fall short of making sense of those quantitative, often more important and more complex, aspects.

Due to these limitations, an extended philosophy of mechanisms, accompanied

by an updated account of mechanistic explanation, is called for (Bechtel and Abrahamsen 2010; Brigandt 2013; Levy and Bechtel 2016). However, although Levy and Bechtel (2016), among others, have pointed out the limitations of the NMP and signposted the direction for an extension, they have not fully fleshed out what that extended philosophy would be. For the moment, let me list those key features, as singled out and agreed upon by these philosophers, that an extended conception of a mechanism must be able to capture. First, the extended framework must treat a mechanism as a non-linear, dynamic complex system that may involve feedback/feedforward loops. Second, in addition to the qualitative thinking, the extended framework must also facilitate quantitative thinking. Third, as a result, the extended philosophy must come even closer to real scientific practice. Given these ingredients, it is time to portray the full image.

4. Mechanism 2.1: A Dynamic Causal Approach

I propose that a mechanism is a *dynamic causal system* that involves various components interacting, typically non-linearly (though sometimes linearly), with one another to produce a phenomenon of interest. In agreement with the NMP, my approach also holds that a mechanism involves four elements: a phenomenon/behavior to be explained, components/parts/entities, interactions/activities/operations, and spatiotemporal organization/structure. Besides, it also considers the multilevel character of mechanisms. However, my approach

differs from the NMP in two important aspects. First, it treats a mechanism as a dynamic system that may involve non-linear interactions and feedback/feedforward loops, and second, it *explicitly* views a mechanism as a *causal structure* composed of components and their causal connections (Here I am not denying that many advocates of the NMP also treat a mechanism as a causal structure. The point is that they only do so implicitly or qualitatively. So, by 'explicitly' I mean a mechanism is *formally* represented as a causal structure using certain quantitative tools, e.g., causal graphs (Spirtes et al. 2000; Pearl 2009).

This approach does not come out of the blue. Rather, it reflects how scientists—especially those neuroscientists—in practice conceptualize a mechanism (Friston et al. 2003, 2009, 2017; Stephan et al. 2007; Rubenstein et al. 2016). To see how this approach can make sense of scientific practice and therefore offer us an extended conception of mechanisms, consider an example drawn from neuroscience. Neuroscientists wonder how human brains respond to stimuli, e.g., visual words. The question they are asking is what mechanism underlies the observed pattern regarding humans' response to visual stimuli. To answer this question, they hypothesize a mechanism involving five components (i.e., areas) in the brain: visual areas V1 and V4, the inferior temporal gyrus (BA37), the angular gyrus (BA39), and the superior temporal gyrus (STG). The hypothesized mechanism is depicted below:

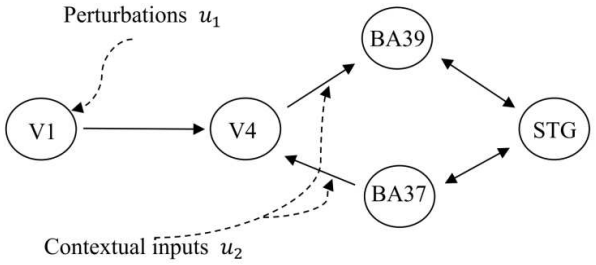


Figure 1. A schematic representation of a neuronal mechanism responsible for bringing about the observed stimuli-response pattern in humans. The figure is adapted from Friston et al. (2003, 1275).

Obviously, this mechanism involves feedback loops. Also, the mechanism can be interpreted as a causal structure, for all the arrows, both the one-way and two-way arrows, denote causal connections.[4] These causal connections are termed *effective connectivity*, denoting "the influence that one neuronal system exerts over another in terms of inducing a response" (Ibid., 1277). As can be seen from the figure, there are two kinds of stimuli/inputs that influence the system: a stimulus can induce a response by either exerting direct influences over a specific region, e.g., u_1, or exerting indirect effects by modulating the coupling (i.e., the causal connection) among regions, e.g., u_2. Attention to a particular feature is a case of the second kind of stimulus/input, for differing degrees of attention usually can result in different strengths of the coupling between the same set of regions. In total, there are three

[4] Notice that this approach differs from the causal graphical theory (Spirtes et al. 2000; Pearl 2009), since it allows cyclic causal structures while the latter does not.

types of interactions: (1) the direct influence of inputs on brain areas, (2) the intrinsic coupling among brain areas, and (3) the modulation of the intrinsic coupling induced by inputs.

We have not yet seen how the mechanism can be dynamic. Given Figure 1, mental simulation may help us roughly understand how the mechanism works, but it offers no help in understanding the mechanism dynamically. To do so, we must be equipped with some mathematical tools. The deterministic differential equations are often the sought-after tools by neuroscientists.[5] Now, we assign a state variable x_i to each region of the mechanism, describing some neurophysiological properties of that region, e.g., postsynaptic potentials. These state variables can interact with one another, namely, one state variable's change relies at least upon (the change of) one other state variable. The set of interactions between the state variables then can be expressed by a set of ordinary differential equations:

$$\frac{dx}{dt} = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{bmatrix} = F(x) \qquad (1)$$

Yet, this set of equations is insufficient to specify the mechanism. To begin with, the set of equations does not give us any information about the specific form, or the nature, of the causal relationships, f_i. Hence, a set of parameters, denoted by θ, that

[5] The other options are state space models, iterative maps, etc.

encodes the information about the form and strength of the causal relationships is required. The set of dependence/causal relationships, however, does define the structure/organization of the mechanism (Stephan et al. 2007, 130). Second, since the mechanism is an open system that exchanges matter, energy and/or information with its environment, the inputs into the system, denoted by the vector function $u(t)$, should also be considered. By expanding equation (1) along these two lines, we obtain a general nonlinear state equation for the system:

$$\frac{dx}{dt} = F(x, u, \theta) \qquad (2)$$

This equation describes how a state variable's change is a function of some neurophysiological influences exerted by some state variables (including itself at an earlier time) and some inputs, and establishes a mapping between the system dynamics and the system structure. It offers

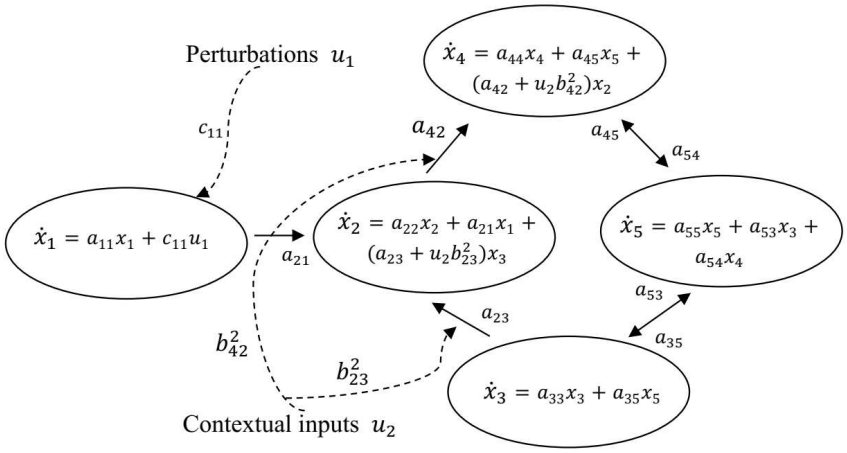
> "A causal description of how system dynamics results from system structure, because it describes (i) when and where external inputs enter the system; and (ii) how the state changes induced by these inputs evolve in time depending on the system's structure. Given a particular temporal sequence of inputs $u(t)$ and an initial state $x(0)$, one obtains a complete description of how the dynamics of the system […] results from its structure […]" (Ibid., 130).

The equation is general because it provides an overarching framework for representing neural systems that can be implemented in different ways. One such an implementation, a bilinear approximation,[6] represents the system dynamics using a bilinear differential equation:

$$\begin{aligned}\frac{dx}{dt} &= F(x, u, \theta) \\ &= Ax + \sum u_j B^j x + Cu \\ &= \left(A + \sum u_j B^j\right)x + Cu \end{aligned} \quad (3)$$

where A is the connectivity matrix denoting the intrinsic coupling among brain areas when no input is present, B^j are the induced connectivity matrices denoting the change of the intrinsic coupling induced by the jth input, and C is the matrix standing for the direct influences of inputs on brain areas. Together, they constitute the parameter set $\theta = \{A, B^j, C\}$ to be estimated. With the parameter set at hand, the mechanism represented in Figure 1 can be redrawn below:

[6] A bilinear approximation is achieved in the following way: the differential equations for each state variable and for each input are linear individually, but nonlinear jointly. For details of this method, see Svoronos et al. (1980).



$$\begin{bmatrix}\dot{x}_1\\ \vdots\\ \dot{x}_5\end{bmatrix} = \left\{\begin{bmatrix} a_{11} & & \cdots & & 0\\ a_{21} & a_{22} & a_{23} & & \\ \vdots & & a_{33} & & a_{35}\\ & a_{42} & & a_{44} & a_{45}\\ 0 & \cdots & a_{53} & a_{54} & a_{55}\end{bmatrix} + u_2\begin{bmatrix}0 & \cdots & 0\\ & b_{23}^2 & \\ \vdots & \ddots & \vdots\\ & b_{42}^2 & \\ 0 & \cdots & 0\end{bmatrix}\right\}\begin{bmatrix}x_1\\ \vdots\\ x_5\end{bmatrix} + \begin{bmatrix}c_{11} & 0\\ \vdots & \vdots\\ 0 & 0\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix}$$

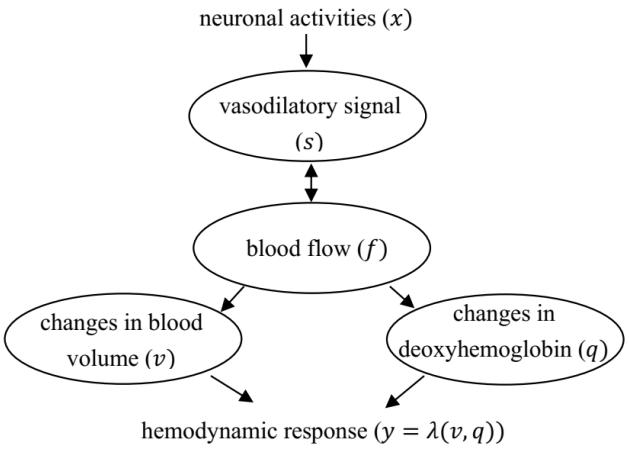
$$\dot{x} = \left(A + \sum_j u_j B^j\right)x + Cu$$

Figure 2. A schema that re-depicts the mechanism in Figure 1 using the differential equations. The lower panel presents the differential equations shown in the upper panel in a matrix form, which can be further simplified using the parameter matrices A, B^j and C. The figure is adapted from Friston et al. (2003, 1279).

In this scenario, each state variable's change, $\dot{x}_i$, is a function of its own state at an

earlier time, at least one other state variable, and some external inputs.

So far, we have shown in detail how a mechanism can be dynamic, and how a mechanism's dynamic character can be properly captured with the help of certain quantitative tools. However, that is not the end of the story. To fully understand a mechanism, it is standard practice that neuroscientists look deeply into each area of the mechanism and treat each as a dynamic system, i.e., a sub-mechanism.[7] More specifically, the sub-mechanism in our example is this: changes in neuronal activity induce a vasodilatory signal which results in changes in blood flow, which in turn cause changes in blood volume and deoxyhemoglobin content. Then, blood volume and deoxyhemoglobin content nonlinearly generate measurable responses of that area. The sub-mechanism of each area is depicted below:



[7] Doing so is partly because each state variable, as representing some neuronal activities, can induce measurable hemodynamic responses, but the causal architecture of the mechanism itself is not observable. So, this is a way to get access to the causal architecture of the mechanism.

Figure 3. A schema that depicts the sub-mechanism of each area of the mechanism. The figure is adapted from Stephan et al. (2007, 133).

This sub-mechanism involves four hemodynamic state variables (s, f, v and q), and a parameter set ϑ. To understand this sub-mechanism dynamically, we, again, need appeal to a set of differential equations that captures the (causal) relationships between these state variables employing the parameter set ϑ.[8] Finally, we obtain a full picture of the mechanism involving two levels (the mechanism-level and the component-level):

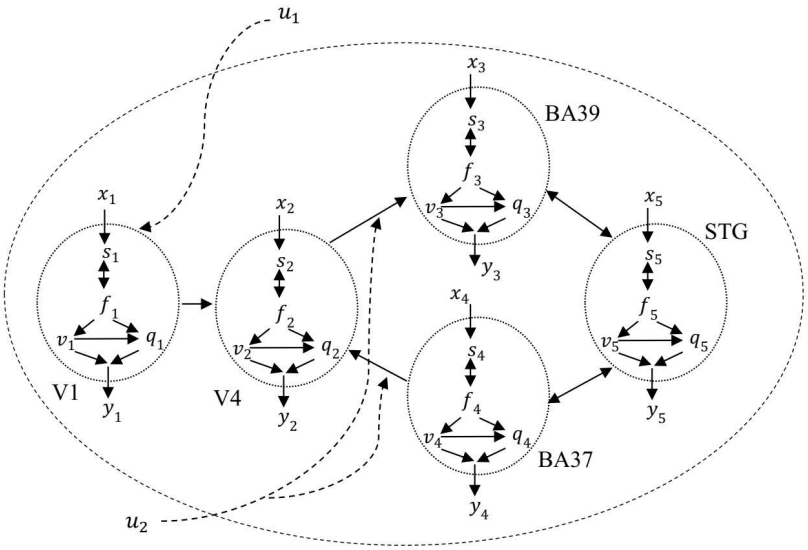


Figure 4. A schema that represents a mechanism and its sub-mechanisms.

[8] This parameter set and the parameter set θ for the system dynamics constitute the whole parameter set $\{\theta, \vartheta\}$, which can be estimated from the measured signal data using a Bayesian estimation approach. The estimation procedure can be found in Friston et al. (2003).

This schematic graph, as depicting a causal structure, together with the quantitative tools necessary to capture the nonlinear, dynamic aspects embodied in the causal structure, constitute the basis for proposing that a mechanism is a dynamic causal system that involves various components interacting, typically non-linearly, with one another to produce a phenomenon of interest.[9] The next section will discuss the key features of this approach, and the philosophical implication it delivers.

5. Discussion

5.1. What is a mechanism, again?

The dynamic causal approach shares with the NMP all those important insights regarding the conception of mechanisms. For example, it agrees that a mechanism consists of four basic elements: a phenomenon to be explained, various components, interactions among these components, and a spatiotemporal organization/structure. Moreover, it treats a mechanism as a multilevel system. Figure 4 in the last section

[9] For the limitations of space, this essay does not fully show how the dynamic, quantitative aspects of the mechanism under consideration are unpacked. For those interested in these details, please see Bechtel and Abrahamsen (2010), where they demonstrate via a similar case, i.e., circadian rhythms, that the dynamic, quantitative aspects can be understood only when certain quantitative tools are employed.

unambiguously reflects this multilevel feature of a mechanism. Furthermore, this approach subscribes to the view that there is no mechanism *simpliciter*, but only a mechanism *for* a particular phenomenon/behavior. In our neuroscientific example discussed above, neuroscientists only singled out five regions of the brain plus their interactions and dismissed all the rest as irrelevant with respect to the stimulus-response pattern in question. Last but not the least, I concur that scientific practice is our best guide to understanding what a mechanism is—that is, we better look at how scientists conceptualize, hypothesize, represent, discover, and entertain mechanisms.

However, a closer look at neuroscientific practice can lead us to some key points overlooked by many new mechanists. First, as some authors have pointed out (Bechtel and Abrahamsen 2010, 2013; Brigandt 2013; Levy and Bechtel 2013, 2016), a mechanism is essentially a dynamic system. Following these authors, I further proposed that a mechanism is a dynamic causal system such that dynamic and causal aspects are a mechanism's defining features. This understanding implies that a qualitative mindset is no longer sufficient to fully understand mechanisms, so that a philosophical conception of mechanisms should be better equipped with a quantitative thinking. Second, many new mechanists emphasize the distinction between entities/parts and activities/interactions. However, an updated philosophy must be able to accommodate the fact that, being a dynamic system, the boundary between entities/parts and activities/interactions may become blurred in some cases. This is the case in our neuroscientific example, where the boundary is clear in the mechanism

involving five regions, but unclear in the sub-mechanisms since their components stand for some quantities that are not clearly entities, e.g., changes in blood flow, changes in blood volume, etc. Though many would think that these quantities are better classified as activities/interactions, the practitioners do not find this classificatory problem worrisome as long as they believe that the state variables denoting them are meaningful and well-defined.

Third, although some philosophers implicitly regard a mechanism as a causal structure, they fail to fully cash out this idea. In my approach, the organization of a mechanism now is explicitly treated as a causal structure that can be quantitatively described using some mathematical tools, e.g., differential equations. The quantitative tools facilitate understanding the nonlinear, dynamic aspects of the causal structure that a qualitative thinking usually stops short of making sense.[10] Also, this dynamic causal approach largely extends the causal graphical theory in characterizing a causal structure, because it allows a causal structure to be cyclic.[11] The causal structure involves both spatial and temporal dimensions, as the spatial dimension is clearly represented by Figure 4 and the temporal dimension is captured by the set of differential equations (in which each region's change is a function of its own earlier

[10] So, the quantitative tools also facilitate understanding the linear aspects if there are such aspects.

[11] Because the variables in the differential equations are somehow time-indexed, e.g., each variable's change is a function of its own state at an earlier time, the problem of circularity does not arise here.

state, at least one other state variable, and perhaps some external inputs).

Unsurprisingly, the dynamic causal approach ramifies into other issues associated with mechanisms, e.g., mechanistic explanation, the way of representing mechanisms, etc.

5.2. An updated account of mechanistic explanation

I follow those new mechanists in holding that a mechanistic explanation is one that uncovers the underlying mechanism of a phenomenon/behavior of interest. But I further add that a mechanistic explanation is a very complicated practice that often—if not at all times—involves the employment of many different epistemic means, e.g., qualitative tools such schematic drawings and verbal descriptions, and quantitative tools such as causal graphs and differential equations, to unpack the dynamic, causal aspects of a mechanism. This view does not deny the value of qualitative tools in offering mechanistic explanation, but it does insist that those qualitative tools can provide explanation only when the explanatory task does not require us to unravel the dynamic aspects of the mechanism.

So, in accordance with Levy and Bechtel (2016), this view regards mechanistic explanation as dynamic in two related senses: on the one hand, the mechanism itself is a complex, dynamic system, and on the other, the process of constructing, articulating and evaluating a mechanistic explanation based on the mechanism in question is also a dynamic matter. This dynamic nature can be reflected by, but not restricted to, the

following scenarios: some parts of a larger system regarded as irrelevant to explaining a phenomenon of interest at an earlier time may be incorporated into a new explanation that treats them as relevant, an explanation may take a different form when a new mathematical tool is invented or when a new component/interaction is identified, a mechanism may at some later stage be embedded into a larger mechanism to explain a phenomenon of interest, etc.

This view also suspects the dichotomy made between mechanistic and mathematical explanation.[12] Some authors maintain that there is a clear-cut boundary between mechanistic and mathematical explanation and that they are competitors rather than comrades (e.g., Craver 2006; Winter 2006). However, our updated account of mechanistic explanation, based on the dynamic causal approach, is able to show that mathematical elements play an indispensable role in building a mechanistic explanation. This is the case in our neuroscientific example, where the set of differential equations is the key to revealing the dynamic aspects of the mechanism. This position goes in tune with many philosophers who either show that mathematical elements are indispensable for a mechanistic explanation (e.g., Bechtel and Abrahamsen 2010, 2013; Brigandt 2013), or demonstrate that constructing mechanistic explanation in the life sciences usually takes an integrative strategy where both mechanistic and mathematical elements figure prominently and work

[12] Mathematical explanation here narrowly means those using mathematics to explain physical phenomena, rather than those *purely mathematical explanations*. See Colyvan (2012) for the distinction.

collaboratively (e.g., Fagan 2012; Boogerd et al. 2013; Green et al. 2015).[13]

5.3. A new way of representing mechanisms

A new conception of mechanisms is usually coupled with a new way of representing mechanisms, and, on the other hand, a new way of representing mechanisms typically reflects a new conception of mechanisms. This two-way dependence relationship has been instantiated in our neuroscientific example, where neuroscientists' conceptualizing mechanisms as dynamic causal systems urges them to appeal to relevant mathematical tools to capture this dynamic causal nature, and the way they represent mechanisms employing these tools also reveals that they think of the mechanisms as dynamic causal systems. Most prominently, they employ differential equations and causal graphs to capture those dynamic causal aspects of a mechanism.

We must note that there might be different ways of representing mechanisms, which may reflect distinct ways of conceptualizing mechanisms. In fact, Casini et al. (2011) and Gebharter and Kaiser (2014) have proposed two alternatives. Casini et al. (2011) attempt to represent a mechanism as a recursive Bayesian network, where each variable at a higher-level can be described as a sub-mechanism at a lower-level. However, though this approach captures the hierarchical and causal nature of

[13] Some also argue that the mathematical elements are part of a broader practice of building mechanistic explanations (Kaplan and Craver 2011; Matthiessen 2017).

mechanisms, it seems unclear how it can treat mechanisms as dynamic systems.[14] Gebharter and Kaiser (2014)'s approach comes closer to my approach, for it respects both the dynamic and causal aspects of mechanisms. But it differs from my approach since it brings the dynamics to the scene via adding time index to each variable, e.g., x_{t_1}, x_{t_2} denote $neuron_x\ firing\ at\ t_1$ and $neuron_x\ firing\ at\ t_2$. This usually results in a very complicated causal structure and therefore seems unpractical.

Notice that this short section is not intended to assess the plausibility/implausibility of different representational strategies, but rather to point out that there are alternatives available and each may have its own merits and shortcomings.

6. Conclusion

Based on neuroscientific practice, I have proposed a dynamic causal approach to characterizing the notion of mechanisms. This approach shares with the NMP all those insights about mechanisms, but also offers an extended, updated conception that highlights the dynamic causal aspects of mechanisms and that comes closer to real scientific practice.

[14] For a more comprehensive criticism, see Gebharter (2014).

References

Bechtel, William. 2006. *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. Cambridge: Cambridge University Press.

Bechtel, William, and Adele Abrahamsen. 2010. "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science." *Studies in History and Philosophy of Science Part A* 41 (3): 321–33.

Bechtel, William, and Robert Richardson. 1993. *Discovering Complexity*. Princeton: Princeton University Press.

Boogerd, Fred, Frank Bruggeman, and Robert Richardson. 2013. "Mechanistic Explanations and Models in Molecular Systems Biology." *Foundations of Science* 18 (4): 725–44.

Brigandt, Ingo. 2013. "Systems Biology and the Integration of Mechanistic Explanation and Mathematical Explanation." *Studies in History and Philosophy of Science Part C* 44 (4): 477–92.

Casini, Lorenzo, Phyllis McKay Illari, Federica Russo, and Jon Williamson. 2011. "Models for Prediction, Explanation and Control: Recursive Bayesian Networks." *Theoria* 26 (1): 5–33.

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

Darden, Lindley. 2006. *Reasoning in Biological Discoveries: Essays on Mechanisms, Interfield Relations, and Anomaly Resolution*. Cambridge: Cambridge University

Press.

Fagan, Melinda Bonnie. 2012. “Waddington Redux: Models and Explanation in Stem Cell and Systems Biology.” *Biology & Philosophy* 27 (2): 179–213.

Friston, Karl, Lee Harrison, and Will Penny. 2003. “Dynamic Causal Modelling.” *Neuroimage* 19 (4): 1273–302.

Friston, Karl. 2009. “Causal Modelling and Brain Connectivity in Functional Magnetic Resonance Imaging.” *PLoS Biology* 7 (2): e1000033.

Gebharter, Alexander. 2014. “A Formal Framework for Representing Mechanisms?” *Philosophy of Science* 81 (1): 138–53.

Gebharter, Alexander, and Marie Kaiser. 2014. “Causal Graphs and Biological Mechanisms.” In *Explanation in the Special Sciences*, edited by Marie Kaiser, Oliver Scholz, Daniel Plenge, and Andreas Huttemann, 55–85. Springer.

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

Green, Sara, Melinda Fagan, and Johannes Jaeger. 2015. “Explanatory Integration Challenges in Evolutionary Systems Biology.” *Biological Theory* 10 (1): 18–35.

Kaplan, David, and William Bechtel. 2011. “Dynamical Models: An Alternative or Complement to Mechanistic Explanations.” *Topics in Cognitive Science* 3 (2011) 438–44.

Kaplan, David, and Carl Craver. 2011. “The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective.” *Philosophy of Science* 78 (4): 601–27.

Levy, Arnon, and William Bechtel. 2013. “Abstraction and the Organization of Mechanisms.” *Philosophy of Science* 80 (2): 241–61.

Levy, Arnon, and William Bechtel. 2016. “Towards Mechanism 2.0: Expanding the Scope of Mechanistic Explanation.” *PhilSci-Archive*.

Machamer, Peter, Lindley Darden, and Carl Craver. 2000. “Thinking about Mechanisms.” *Philosophy of Science* 67 (1): 1–25.

Matthiessen, Dana. 2017. “Mechanistic Explanation in Systems Biology: Cellular Networks.” *The British Journal for the Philosophy of Science* 68 (1): 1–25.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Rubenstein, Paul, Stephan Bongers, Bernhard Schölkopf, and Joris Mooij. 2016. “From Deterministic ODEs to Dynamic Structural Causal Models.” *ArXiv Preprint ArXiv:1608.08028*.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Cambridge: MIT press.

Stephan, Klaas, Lee Harrison, Stefan Kiebel, Olivier David, Will Penny, and Karl Friston. 2007. “Dynamic Causal Models of Neural System Dynamics: Current State and Future Extensions.” *Journal of Biosciences* 32 (1): 129–44.

Svoronos, Spyros, George Stephanopoulos, and Rutherford Aris. 1980. “Bilinear Approximation of General Non-Linear Dynamic Systems with Linear Inputs.” *International Journal of Control* 31 (1): 109–26.

Paper to be presented at PSA2020: The 27th Biennial Meeting of the Philosophy of Science Association, November 19-22, 2020, Baltimore, MD.

"It takes two to make a thing go right": The coevolution of technological and mathematical tools in neuroscience

Luis H. Favela[1, 2]

[1] Department of Philosophy, University of Central Florida

[2] Cognitive Sciences Program, University of Central Florida

Author Note

Luis H. Favela https://orcid.org/0000-0002-6434-959X

Correspondence concerning this article should be addressed to Luis H. Favela, Department of Philosophy, University of Central Florida, 4111 Pictor Lane, Suite 220, Orlando, FL 32816-1352. E-mail: luis.favela@ucf.edu

"It takes two to make a thing go right": The coevolution of technological and mathematical tools in neuroscience

Abstract

Some philosophers of neuroscience have recently argued that the history of neuroscience is principally a history of technological tool development. Across these claims, there is little to no mention of data analysis methods nor their underlying assumptions. Here, I argue that mathematical tools have played crucial roles in the history of neuroscience. First, I present the Hodgkin-Huxley model as an example of research constrained by technological limitations and mathematical assumptions. Second, I highlight scale-free neuronal dynamics and explain how that discovery required both technological and mathematical advancements. I conclude by discussing consequences for explanations in neuroscience.

Keywords: Hodgkin-Huxley model, mechanism, neuronal dynamics, scale-free, tool development

Word count

4,470 (including abstract 96 words)

~4,970 (if 100 words added per figure; including abstract 96 words)

2

It takes two to make a thing go right.

—Rob Base (Ginyard) and DJ E-Z Rock (Bryce), *It takes two*

1. Introduction

There should be no doubt that technological developments have played significant roles throughout the history of scientific discoveries and progress. This is as true in the physical sciences (e.g., particle accelerators in physics) as in the life sciences (e.g., microscopes in biology). What is less apparent is the role mathematical developments have played in facilitating and supporting many of those discoveries. Mathematical tools for analyzing data may not be at the forefront of discoveries centering on the physical structure of investigative targets of interest (e.g., cells); but they certainly are crucial in research focused on the dynamics of phenomena (e.g., planetary motion). In short, for science to progress, research on the movement and temporal aspects of phenomena often require the coevolution of technological *and* mathematical tools.

Recently, it has been increasingly argued by some philosophers of neuroscience that experimental tools are not just important but are fundamental to neuroscience research (e.g., Bickle, 2016). Put in its most extreme terms, the line of thought goes like this: From Golgi's staining technique to functional magnetic resonance imaging, and from deep brain stimulation to optogenetics, the history of neuroscience is principally a history of tool development. Moreover, it has been argued that this history is best characterized as one that exhibits reductionist (Bickle, 2006, 2016) and mechanistic explanations (Craver, 2002, 2005). Across

these claims, little to no mention of data analysis methods are mentioned nor the underlying assumptions of those methods. Here, I argue that the mathematical assumptions of applied data analyses have played crucial roles in the history of neuroscience. First, I present the Hodgkin and Huxley model of action potentials as an example of research constrained by technological and mathematical limitations of the time. Second, I draw attention to a feature of neurons that is overlooked by the Hodgkin-Huxley model: scale-free dynamics. After describing scale-free dynamics, I then point out a consequence scale-free neuronal dynamics has for mechanistic explanations of neuronal activity. I conclude by discussing the necessity of mathematical developments in providing appropriate accounts of scale-free neuronal activity.

2. Hodgkin-Huxley model and scale-free neuronal dynamics

The canonical Hodgkin and Huxley (1952) model of action potentials in the squid giant axon is considered "the single most successful quantitative model in neuroscience" (Koch, 1999, p. 171). The majority of the details of the model are not essential for my current aims. For detailed explanations of this model see Gerstner, Kistler, Naud, and Paninski (2014), as well as Koch (1999) for discussion and further references. For now, it is important to understand that this model treats the action potential as an event that is "all-or-none" in that it occurs within distinctly defined timescales (e.g., ; Bear et al., 2016; Figure 1). Moreover, those timescales have a lower boundary, specifically, 10 milliseconds (ms) in the canonical Hodgkin-Huxley model (Hodgkin & Huxley, 1952, p. 528; Koch, 1999,

p. 334; Marom, 2010, p. 23). What that means is that the action potential of a neuron (i.e., its "spike" of activity) is treated within the Hodgkin-Huxley model as occurring at least 10 ms from initiation to termination of all involved processes (Marom, 2010, p. 22).

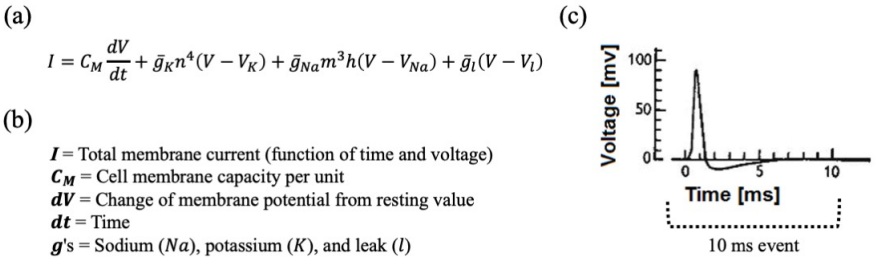


Figure 1. Hodgkin-Huxley model. (a) The canonical Hodgkin and Huxley (1952) model of action potentials in the squid giant axon. (b) Definitions of key model variables. (c) The basic shape of an action potential as produced by Hodgkin-Huxley model. (Modified and reprinted with permission from Wikipedia. CC BY-SA 4.0.) The x-axis captures the entire range of time in which an action potential occurs. According to the model, the lower temporal boundary of an action potential is 10 ms. This means that the entire event, from start to finish, occurs within that time frame.

As is well-known (e.g., Marom, 2010), although there were empirically justifiable reasons at the time (e.g., Adrian & Zotterman, 1926), defining the "action potential" as a 10 ms event was due to investigator observational preferences in combination with technological limitations. Observational preferences were constrained by the limits of the recording technology, namely,

the voltage clamp. Although the voltage clamp was instrumental in providing the data that lead to the development of the Hodgkin-Huxley model, it was limited in its ability to record the full range of ion channels, charged particles, and other physiologically relevant features of neuronal activity (Schwiening, 2012). This resulted in the need to sum across molecular activity (Gerstner et al., 2014)—certainly a necessity when calculating at the molecular scale—and collapse other physiological features into imprecise "leak" terms, a sort of "catch all" variable used in models that have causally relevant features that have not been precisely measured. Other limitations involved the manner in which the data was calculated. Hodgkin and Huxley calculated data from the voltage clamp via hand calculators (Koch, 1999, p. 160). Specifically, Hodgkin and Huxley utilized a mechanical calculator, the Brunsviga 20 (Figure 2), which required them to spend a few weeks and many thousands of rotations of the mechanical calculator's crank (Schwiening, 2012).

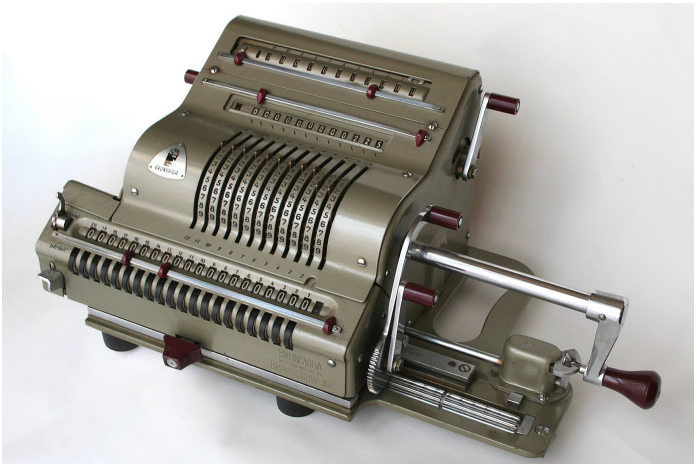


Figure 2. The Brunsviga 20, "one of the most popular mechanical calculators. It was produced up to the early 1970s and marketed with the slogan 'Brains of

Steel'" (Schwiening, 2012). (Reprinted with permission from Wikipedia. CC BY-SA 2.0 DE).

Although the canonical Hodgkin-Huxley model is described by some as being linear in nature (e.g., Gerstner et al., 2014; Hodgkin & Huxley, 1952, pp. 538-540), there is debate about whether or not it is able to capture the relevant types of nonlinearities exhibited by feedback that are now established as occurring during action potentials (e.g., Marom, 2010; Schwiening, 2012). Regardless whether or not the canonical Hodgkin-Huxley model is linear or nonlinear, or can capture particular forms of feedback, it is clear now that even single neurons are appropriately understood as nonlinear systems (e.g., Izhikevich, 2007).

Advancements in recording technologies have facilitated the ability of neuroscientists to obtain more detailed data on neuronal activity (e.g., multielectrode arrays; Gross, 2011), making it possible to record more detailed and accurate data from longer timescales of neuron activity. As a result, it is becoming increasingly evident that the relevant timescales for explaining even "basic" single-neuron activity requires looking below and above that 10 ms window. Action potentials do not appear to have strictly defined windows of activity, specifically, nonlinearities in the forms of feedback and hysteresis significantly contribute to the event. Instead of viewing action potentials as having clear startup and finish conditions (Figure 1), it is more accurate to view action potentials as continuous, nonlinear cycles. This is clearly depicted in early

models, such as the FitzHugh-Nagumo model (FitzHugh, 1961; Nagumo et al., 1962; Figure 3a) and more recent models, such as the Izhikevich model (Izhikevich, 2007; Figure 3b).

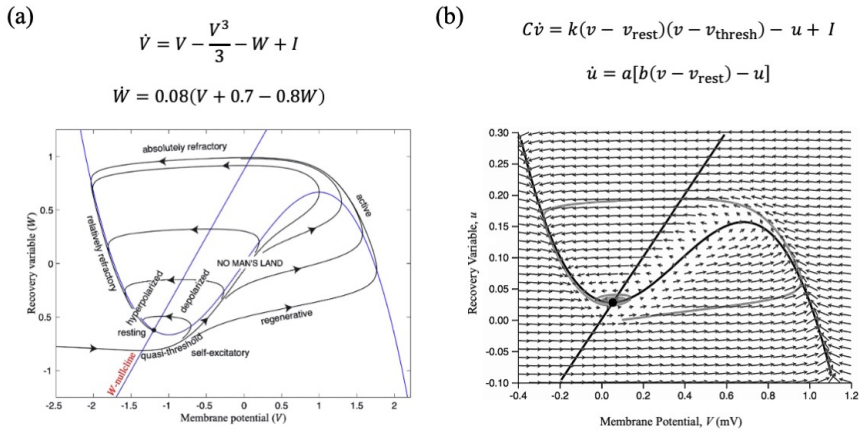


Figure 3. Models of single-neuron activity. (a) FitzHugh-Nagumo model and phase space portrait. (Modified and reprinted with permission from Scholarpedia. CC BY-NC-SA 3.0.) (b) Izhikevich model and phase space portrait. (Modified with permission from J. Terwilliger, 2018.)

As mentioned above, there is debate as to the degree or not that the canonical Hodgkin-Huxley model accounts for a wide range of nonlinear features of action potentials, such as hysteresis. I am not entering that debate here. Instead, I focus on a particularly notable recent finding that has resulted from improved recording technologies. That finding is the apparently scale-free nature of neuronal activity. At its most general, a phenomenon is "scale-free" (or "scale invariant") when its structure (i.e., behavioral, spatial, and/or temporal) is statistically self-similar from various points of observation (Bak, 1996; Gisiger,

2001). Many illustrative examples of spatial scale-free structures are found in fractal geometry (Mandelbrot, 1983; Figure 4).

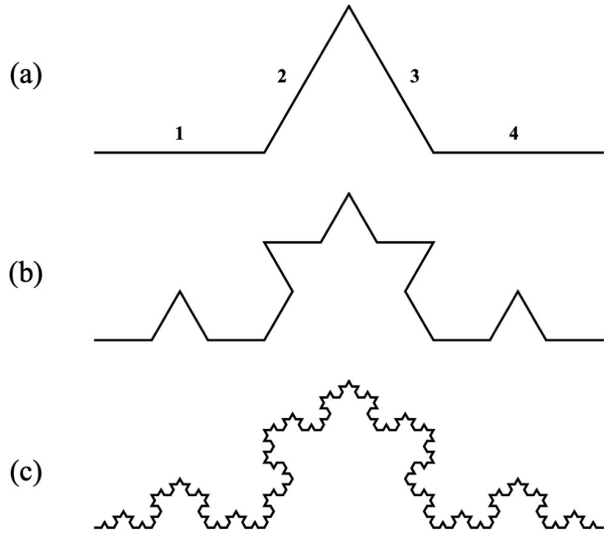


Figure 4. The Koch triangle is an example of a spatial fractal. Here, three iterations of self-similarity are depicted (a,b,c). (Modified and reprinted with permission from Wikipedia. CC BY-SA 3.0.)

Scale-free properties have become particularly popular in recent years in regard to network structure, where few nodes have many connections and many nodes have few connections. Consequently, such networks have no specific or average number of connections that characterize the entire system. Mathematically speaking, scale free structures can be characterized by their power-law distribution (He, 2014). It has become commonly accepted that many phenomena and systems of diverse composition are scale free in this way, for

example, cellular metabolism, Hollywood actors that have worked together, sexual relationships, and the World Wide Web (Barabasi & Bonabeau, 2003). There is increasing evidence that neural systems exhibit many scale-free properties (e.g., Boonstra et al., 2013; He, 2014). These properties are exhibited from neuronal network connections to neuron branching patterns. For current purposes, I focus on the scale-free dynamics exhibited by neuronal dynamics (for a wide range of examples see Boonstra et al., 2013). In short, neuronal dynamics are considered "scale free" when there is no single time scale that properly characterizes its activity, which includes attempting to define an event as occurring within specific windows of time. There are a number of consequences that result from the fact that many neural systems exhibit scale-free spatial or temporal structure. In the next section I explore one such consequence, specifically, the inability of mechanistic explanations to account for scale-free neuronal dynamics.

3. Consequences of scale-free dynamics for explanations in neuroscience

In a recent paper, Bechtel (2015) argues against the claim that scale-free biological phenomena cannot be explained mechanistically. He rejects the following argument, which I summarize as follows:

1. Mechanistic explanations require that the phenomena being explained have well-defined boundaries, such as a temporal boundary.
2. Many biological phenomena exhibit scale-free features.
3. Scale-free phenomena have no well-defined temporal boundaries.

4. Therefore, scale-free biological phenomena cannot be explained mechanistically.

Marom (2010) presents such an argument and serves as one of Bechtel's targets. Marom argues that there is empirical evidence suggesting that neuronal activity is scale-free and, thus, is just the type of biological phenomenon that cannot be explained mechanistically. Marom's argument includes discussion of the Hodgkin-Huxley model, which leads him to conclude:

> Indeed, the lesson from our journey across levels of organization, from behavior through neural assemblies to single neurons and proteins, suggests that dreams on all-encompassing microscopic timescale-based descriptions, aimed at explaining the temporal richness of macroscopic levels, should be abandoned. Other approaches are called for. (2010, p. 23)

In short, Marom claims that there are no uniquely defined timescales that could justify defining action potentials as events that have a lower boundary of 10 ms. Consequently, macroscale neuronal activity that appear scale-free are not merely the result of additive or linear combinations of microscale contributions. Instead, they are truly scale-free: the micro timescales contribute to and constrain the macro timescales, but so too does the macro contribute to and constrain the micro, such that no single scale serves a more fundamental explanatory role than the others.

Bechtel's reply to Marom is that scale-free phenomena can still be explained mechanistically. But to do so requires that we appreciate the role of

mechanisms in scientific practice. According to Bechtel, scientists often posit "bounded mechanisms" for the purposes of testing hypotheses (2015, pp. 84-85). A scientist can understand that a phenomenon is interconnected (e.g., networks) and still pursue a mechanistic account of that phenomenon by drawing boundaries around that organism. Those bounded mechanisms are not abstractions, however. "Abstractions," according to Bechtel, leave information out. Instead, those bounded mechanisms are idealizations. Idealizations, according to Bechtel, are models with *simplifying* falsehoods (2015, p. 85). For example, if phenomenon X is understood to be highly interconnected, an explanation of X that assumes that it is not affected by all of those connections would be an abstraction. But to localize X to, for example, its nearest neighbors, is to provide a "*first approximation*" (2015, p. 85; italics in original) that appreciates the practical challenges of accounting for all the actual connections. Such an explanation would be both an idealization and a mechanism.

Although he accepts that neuronal dynamics can be scale-free, Bechtel remains committed to providing mechanistic explanations of those dynamics. Accordingly, Bechtel remains committed to mechanisms being bounded, on the further stipulation that such bounded mechanisms are idealizations and not abstractions. For example, the action potential is a "bounded mechanism" that occurs within 10 ms windows. Such an idealization is acceptable because it makes the timescales of that phenomenon tractable to investigators' cognitive limitations (2015, p. 92). Thus, the Hodgkin-Huxley model can be understood as an idealization of action potentials, with the 10 ms feature being a simplifying

falsehood—though not an abstraction that leaves out relevant features. This is a very streamlined presentation of Bechtel's argument, for example, he makes a further claim that such idealized mechanisms can point out areas for further investigation in a mechanistic explanation. What matters for my current purposes, is Bechtel's attempt to make room within mechanistic accounts to explain scale-free activity.

There is a lot in Bechtel's reply to Marom to agree with, for example, the fact that scientists are epistemically-limited creatures who need to simplify some phenomena in order to get an intellectual grip on them. However, I think Bechtel's reply overlooks a central issue raised by Marom. If mechanisms are, by definition, *bounded*, then scale-free phenomena (e.g., scale-invariant, fractal, flicker noise, power laws, etc.; Gisiger, 2001) are, by definition, not mechanisms. In the case of action potentials, the canonical Hodgkin-Huxley model sets a lower boundary on the phenomenon at 10 ms. In other words, it treats action potentials as starting and finishing within windows of time of at least 10 ms (Figure 1c). As discussed above, such a claim was justified as being consistent with the best science of the time (e.g., Adrian & Zotterman, 1926). With that said, it was constrained by technological (voltage clamp) and mathematical (the type of calculations that could be conducted on a Brunsviga 20 calculator; Figure 2) limitations. Technological advancements have certainly played a role in revealing scale-free dynamics (e.g., multielectrode arrays; Gross, 2011). However, data from advanced equipment alone has not justified the existence of scale-free dynamics in neuronal systems. The other part needed for the right account—

remember, “it takes two to make a thing go right”—is the pairing of data from suitable technology with the appropriate mathematical tools.

In the case of single-neuron activity, the right mathematical tools are those from nonlinear dynamical systems theory (NDST; e.g., Izhikevich, 2007; Liebovitch & Toth, 1990). NDST methods are crucial to assessing scale-free structure, and can contribute to establishing whether a phenomenon is truly scale-free or not and, if so, what kind of scale-free characteristics it has. What’s more, applying NDST methods to complex and nonlinear phenomena typically requires powerful computers. For example, generating phase portraits of relatively simple two-dimensional dynamical systems was often not practical before computers. Hodgkin and Huxley’s “Brains of Steel” mechanical calculator was certainly not up to the task. Thus, the Izhikevich model of single-neuron activity required both the appropriate processing power (i.e., modern computers) *and* data analysis methods (i.e., NDST) in order to provide qualitative and quantitative accounts of that phenomenon’s nonlinear dynamics.

As mentioned above, nonlinear dynamics are not central to my current aims; but scale-free dynamics are. Scale-free properties are a particularly unique set of phenomena in regard to the need for coevolved technological and mathematical tools. Many aspects of mammalian biological phenomena alone exhibit scale-free structures, such as, bronchial tube branching, eye saccades, heart beats, neuronal networks, and postural sway. Accordingly, different mathematical tools are needed to properly determine the ways they are scale-free. For example, detrended fluctuation analysis (Peng et al., 1994) can assess

structural self-similarity in a signal, but will not necessarily make clear if the structure results from linear or nonlinear processes (Bryce & Sprague, 2012). In the case of appropriate mathematical methods for assessing the scale-free dynamics of action potentials, if such activity is, for example, fractal, then it would not have been possible to accurately analyze such data, regardless of technological advancements, until the 1980s. The reason is because the concept "fractals" was not introduced to the broader scientific community until then (Mandelbrot, 1983).

In order to identify fractal scale-free structures, whether resulting from linear or nonlinear processes, the concept "fractals" and their measurement must be part of an investigators toolbox. Fractals, such as the Koch triangle (Figure 4) are paradigm examples of scale invariance: the overall structure of the system is maintained at each level of observation. Such phenomena are thus not appropriately explained in terms that, for example, treat them as having an average value. Instead, as Mandelbrot pointed out, such phenomena are appropriately characterized via a fractal dimension. The fractal dimension provides a quantitative means of characterizing a scale-free phenomenon that accounts for all of its scales. The equation for calculating the fractal dimension is:

$$n = {}^{1}/_{S^{d}}$$

Let's go back to the Koch triangle. For demonstration purposes, we will look at a four-lined Koch triangle (Figure 4). Here *n* is the number of line segments at a particular scale of observation; in this case, it is 4. Next, *S* is the scale factor, or the size reduction at each iteration; here it is 1/3. Our equation is

now: $4 = 1/(1/3)^d$, or $4 = 3^d$. We want to figure out d, or the fractal dimension. To do so, we take the log of both sides: $d = \log 4/\log 3$, which gives us a fractal dimension $d = 1.26$. In English, this means that the fractal dimension of the Koch triangle is 1.26, which means it is not a straight line (1) or a square (2), but closer to being a straight line than a square (1.26). There are various other methods for mathematically assessing fractals and multifractals (Lopes & Betrouni, 2009).

The point of this example is to demonstrate that before Mandelbrot's invention (discover?) of fractal geometry, it was not possible to accurately account for such phenomena, for example, collapsing scale-invariant structures into single values (e.g., arithmetic mean). The consequence for neuronal activity is that it was not until the 1990s (e.g., Liebovitch & Toth, 1990) that scale-free dynamics could be properly identified. Before then, such properties were misidentified via other statistical methods. Since scale-free structures have no primary scale or average scale, they have no specific window to identify as the start and finish boundary. Such a view of neuronal activity is further evidenced by other NDST-based work, such as the Izhikevich model (2007; Figure 3b), which treats action potentials as continuous cycles and not "all-or-none" (cf. Figure 1c). If true, that is, if action potentials are not bounded within discrete windows of time, then action potentials cannot be accounted for mechanistically.

In concluding this section, an important clarification needs to be made in order to address a significant critique of the current line of thought. The critique centers on the notion of "bounded" in regard to natural phenomena. As discussed above, the currently-relevant aspect of the Bechtel/Marom debate centers on the

idea that mechanistic explanations treat targets of investigation as bounded, namely, as having delineated borders, which can be spatial or temporal. The Hodgkin-Huxley model of action potentials and its 10 ms event window were presented as an example of such a bounded mechanism. Scale-free neuronal dynamics was presented as an unbounded phenomenon, which means it is not a phenomenon accessible to mechanistic explanation (i.e., if "mechanistic explanations" include the stipulation of boundedness; see Bechtel, 2015 and Marom, 2010). The critique of this line of thought centers on the point that even scale-free neuronal dynamics are "bounded" in a number of ways, for example, there *is* a window of time in which they occur (e.g., they do not last for months, years, or centuries) and they *are* spatially confined (e.g., they occur in an area of the brain, and not across the whole brain, let alone body). This is a compelling critique. However, it does not address the way in which scale-free dynamics are "unbounded." The way in which scale-free dynamics are unbounded concerns the inability of single, bounded values to *characterize* the phenomenon. A time series (Figure 5) need not be infinite nor recorded from an event that has no spatial location in order to be scale free. A scale-free time series exhibits the same pattern among windows of various lengths of time. For example, if a heartbeat shows a pattern of activity over 60 minutes, then, to be considered scale-free, that same pattern should be shown in each of two 30 minute windows of time, at each of four 15 minute windows, and so on. In that way, the time series is not properly understood as "bounded" in that there is no single length of time that characterizes the entire signal. That is to say, it is not correct to treat the event as a

bounded 60 minute event, or a 30 minute event, and so on; but in terms of the structure of the patterns across various scales. It is in that sense that Marom argues that neuronal dynamics do not have timescales, and it is in that sense that they are unbounded, and, thus, not properly explained mechanistically.

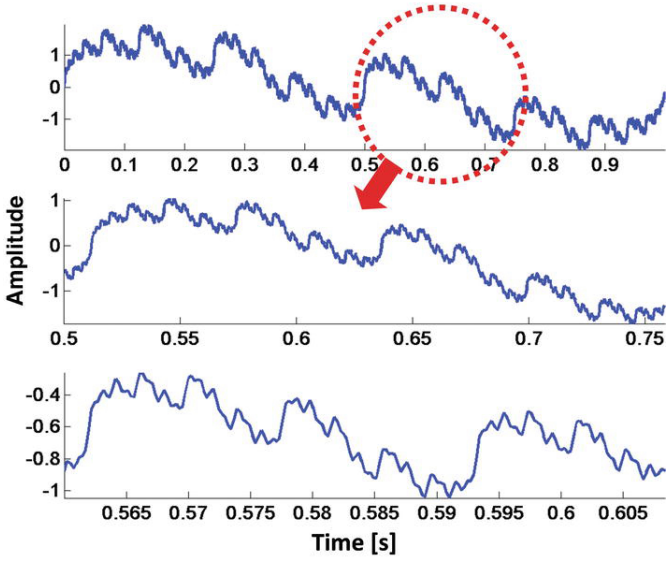


Figure 5. Fractal time series exhibiting scale-free structure at various windows of time. (Reproduced with permission from Armentano et al., 2017. CC BY 3.0.)

4. Conclusion

It is highly unlikely to find disagreement among the scientific research community at large that technological advancements have paved the way for some of the greatest advances and discoveries. What is less often acknowledged—especially in neuroscience—is the necessity of coevolving our mathematical tools with technological advances, and vice versa. Consequently, technological advancements that produce more detailed and accurate data

recording will not alone necessarily provide proper explanations of biological phenomena. Mathematical tools like those provided by NDST are needed as well in order to properly characterize data. The Hodgkin-Huxley model was informed and constrained by the available technological (i.e., voltage clamp) and mathematical (i.e., Brunsviga 20 calculator) tools of the time. Since then, more advanced technology (e.g., multielectrode arrays) and mathematics (e.g., fractal analysis) have highlighted some of the shortcoming of the Hodgkin-Huxley model as a comprehensive model of action potentials across temporal scales. Scale-free neuronal activity provides a rich example of this. In order to identify scale-free activity, researchers needed more accurate measurements, data analyses, and—in this case—new concepts altogether. In order to properly account for scale-free activity, a new concept—namely, fractals and the fractal dimension—was needed, as was accompanying innovative mathematical analyses. One consequence of the existence of scale-free neuronal activity discussed here involves the limitations of mechanistic explanations to account for phenomena that are without discrete temporal boundaries. In sum, an attempt has been made here to demonstrate that it takes two to make progress in neuroscience, namely, both technological and mathematical advancements.

References

Adrian, E. D., & Zotterman, Y. (1926). The impulses produced by sensory nerve endings. Part 3. Impulses set up by touch and pressure. *The Journal of Physiology*, *61*(4), 465-483.

Armentano, R. L., Legnani, W., & Cymberknop, L. J. (2017). Fractal analysis of cardiovascular signals empowering the bioengineering knowledge. In F. Brambila (Ed.), *Fractal analysis - Applications in health sciences and social sciences*. IntechOpen. doi:10.5772/67784

Bak, P. (1996). *How nature works: The science of self-organized criticality*. New York, NY: Springer-Verlag.

Barabasi, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American*, *288*(5), 60-69.

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2016). *Neuroscience: Exploring the brain* (4th ed.). New York, NY: Wolters Kluwer.

Bechtel, W. (2015). Can mechanistic explanation be reconciled with scale-free constitution and dynamics? *Studies in History and Philosophy of Biological and Biomedical Sciences*, *53*, 84-93.

Bickle, J. (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese*, *151*, 411-434.

Bickle, J. (2016). Revolutions in neuroscience: Tool development. *Frontiers in Systems Neuroscience*, *10*(24). doi:10.3389/fnsys.2016.00024

Boonstra, T. W., He, B. J., & Daffertshofer, A. (2013). Scale-free dynamics and critical phenomena in cortical activity. *Frontiers in Physiology: Fractal and Network Physiology*, *4*(79). doi: 10.3389/fphys.2013.00079

Bryce, R. M., & Sprague, K. B. (2012). Revisiting detrended fluctuation analysis. *Scientific Reports*, *2*(315), 1-6. doi:10.1038/srep00315

Craver, C. F. (2002). Interlevel experiments and multilevel mechanisms in the neuroscience of memory. *Philosophy of Science*, *69*(S3), S83-S97.

Craver, C. F. (2005). Beyond reduction: Mechanisms, multifield integration and the unity of neuroscience. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 373-395.

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, *1*(6), 445-466.

Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge, UK: Cambridge University Press.

Ginyard, R. (1988). It takes two [Recorded by R. Ginyard (Rob Base) and R. Bryce (DJ E-Z Rock)]. On *It takes two* [Vinyl]. United States: Profile Records.

Gisiger, T. (2001). Scale invariance in biology: Coincidence or footprint of a universal mechanism? *Biological Reviews*, *76*, 161-209.

Gross, G. W. (2011). Multielectrode arrays. *Scholarpedia*, *6*(3):5749. doi:10.4249/scholarpedia.5749

He, B. J. (2014). Scale-free brain activity: Past, present, and future. *Trends in Cognitive Sciences*, *18*(9), 480-487.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, *117*(4), 500-544.

Izhikevich, E. (2007). *Dynamical systems in neuroscience: The geometry of excitability and bursting*. Cambridge, MA: MIT Press.

Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. New York, NY: Oxford University Press.

Liebovitch, L. S., & Toth, T. I. (1990). Using fractals to understand the opening and closing of ion channels. *Annals of Biomedical Engineering*, *18*, 177-194.

Lopes, R., & Betrouni, N. (2009). Fractal and multifractal analysis: A review. *Medical Image Analysis*, *13*(4), 634-649.

Mandelbrot, B. B. (1983). *The fractal geometry of nature*. New York, NY: W. H. Freeman and Company.

Marom, S. (2010). Neural timescales or lack thereof. *Progress in Neurobiology*, *90*, 16-28.

Nagumo, J., Arimoto, S., & Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, *50*(10), 2061-2070.

Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., & Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Physical Review E*, *49*(2), 1685-1689.

Schwiening, C. J. (2012). A brief historical perspective: Hodgkin and Huxley. *The Journal of Physiology*, *590*(11), 2571-2575.

Causation and the Problem of Disagreement

Enno Fischer

Abstract

This paper presents a new argument for incorporating a distinction between default and deviant values into the formalism of causal models. The argument is based on considerations about how causal reasoners should represent disagreement over causes and it is defended against an objection that has been raised against earlier arguments for defaults.

A number of authors have argued for incorporating a distinction between default and deviant states into the formalism of causal models.[1] A central motivation for this has been the Problem of Isomorphism (Hall (2007); Halpern and Hitchcock (2015)). This problem arises from pairs of target systems that supposedly have isomorphic causal models but give rise to different judgements of actual causation. The idea is that the different judgements are explained by assumptions about particular values that variables typically or normally take on. These assumptions are taken to be captured by the default/deviant distinction.

However, more recently Blanchard and Schaffer (2017) have argued that key instances of the problem can be solved by revising one of the involved models such that it gives a more appropriate representation of the corresponding target system. They also suggest a generalization of this strategy, which I shall call the *adjust-the-model argument*. They argue that, when confronted with an instance of the Problem of Isomorphism, we should suspect that at least one of the involved models is not an appropriate representation of its target system. They also argue that defaults "come close to a free parameter in an otherwise so precise and objectively constrained formalism, which basically gives the theorist leeway to hand-write the result she wants" (192). Thus, according to them, the default/deviant distinction does more damage than good to the formalism of causal models.

In this paper I shall provide a more nuanced account of the benefits of the default/deviant distinction. I shall grant that Blanchard and Schaffer's criticism of defaults as a solution to the Problem of Isomorphism is right. However, there is another problem that is far less prominent: the Problem of Disagreement. I will show that this problem gives rise to a genuinely new argument for incorporating the default/deviant distinction.

The Problem of Disagreement has been introduced by Halpern and Hitchcock (2015). It arises from cases where agents disagree in their causal judgement even though they make the same assumptions about the underlying causal model. The Problem of Isomorphism is related to well-known examples of disagreement over what is 'the

[1] In the following the term 'causal model' will refer to standard causal models (models without defaults), as introduced by Pearl (2000). Models with defaults will be called 'extended causal models' as introduced by Halpern and Hitchcock (2015).

cause' of a given effect as discussed, for example, by van Fraassen (1980). The main difference is that the Problem of Disagreement involves the explicit assumption that the disagreeing agents base their causal claims on the same underlying causal model. Halpern and Hitchcock take this to indicate that the agents' causal judgements depend not only on assumptions about causal structure but also on a distinction between default and deviant behaviour.

I will show that this argument allows two readings. First, it can be read as involving descriptive claims about how agents *do* reason about causal models in contexts where they disagree. This reading seems to be vulnerable to a version of Blanchard and Schaffer's adjust-the-model argument. If two agents disagree about judgements of actual causation, we should expect that these agents also disagree about the underlying causal model. Second, the argument can be read as involving prescriptive claims about how agents *should* reason about causes when they disagree. Here the adjust-the-model argument does not apply. I will argue that it would be wrong to require that the agents support their conflicting causal judgements with different models. Instead, I will argue, causal models should be understood as a representative tool that helps express causal claims that go beyond causal judgements that are based on potentially idiosyncratic normative presumptions. If understood in this way, they can help resolve disagreement over causes by giving a framework for disentangling normative and epistemic dimensions of disagreement. And this function can only be fulfilled if models incorporate the default/deviant distinction. I will illustrate this claim with an example that concerns the causal role of Search and Rescue missions in the Central Mediterranen with regard to increasing numbers of deaths through shipwreck in 2015 and 2016.

In section 1 I introduce Blanchard and Schaffer's arguments against defaults. In section 2 I introduce the Problem of Disagreement and I point out that Halpern and Hitchcock's way of employing it as an argument for defaults is vulnerable to a version of Blanchard and Schaffer's criticism. In section 3 I raise the question what the function of extended causal models should be in instances of disagreement. Based on the example of Search and Rescue Missions (section 4) I argue that defaults help us clarify disagreement over causes (section 5).

1 The Adjust-the-Model Argument

Blanchard and Schaffer put forward three main lines of criticism against incorporating the default/deviant distinction. First, the default/deviant distinction is unnecessary. Instances of the Problem of Isomorphism arise only because one of the involved models does not provide an appropriate representation of the underlying target system. Instead of incorporating the default/deviant distinction, we should adjust the models, adhering to generally accepted aptness constraints. These aptness constraints are rules for selecting a set of variables $\mathcal{V}$ that constitutes the causal model and in the following we shall focus on the rule that "variables should not be allotted values that we are not willing to take seriously" (182). Blanchard and Schaffer take this aptness constraint to help us deal with cases like the gardener/queen example: some flowers would not have died if either the gardener or the Queen of England had watered them and it needs to be explained why we tend to identify only the gardener as an actual cause.[2]

> "It is because we are willing to indulge in the fantasy of the gardener watering the flowers [...], but just can't imagine the queen stooping to the job, that we feel an asymmetry. If so then [the constraint to represent only serious possibilities]—which does independent work—was all we needed to explain the gardener/queen asymmetry. There is no apt causal model in which wiggling whether the queen waters the flowers wiggles the fate of the flowers, because there is no apt causal model that considers so ridiculous a scenario as the queen of England popping by, watering can in hand, to engage in random acts of gardening" (197).

Figure 1A gives a representation of the gardener/queen case that Blanchard and Schaffer consider to be problematic. They argue that this is not an apt model because $Q = 1$ represents a scenario that we are not willing to take seriously. Thus, they suggest eliminating variable Q, which leads to the simpler model in figure 1B, which reproduces the plausible verdict that only the gardener is an actual cause of the flowers' death.

[2] In the gardener/queen case the problem arises from a symmetry that is internal to the model, not from two causal models that have isomorphic structure.

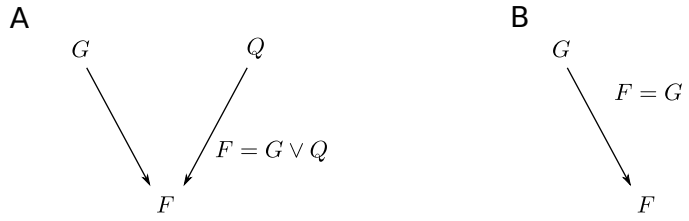


Figure 1: The flowers survive if they are watered by the gardener or by the queen.

Second, Blanchard and Schaffer argue that the default/deviant distinction involves unclarities. Most proponents of defaults relate them to an underlying theory of typicality or normality that involves a range of possibly conflicting standards. Blanchard and Schaffer worry that the unclarity associated with these notions spoils the otherwise precise theoretical framework of causal models.

Third, Blanchard and Schaffer criticise that incorporating the default/deviant distinction is psychologically implausible. Proponents of defaults assume that the judgements evoked by thought experiments like the gardener/queen case reflect judgements that arise from the competent use of a norm-laden notion of actual causation. Instead, according to Blanchard and Schaffer, the fact that causal reasoners ascribe a higher relevance to norm-violating factors or agents is to be explained by norm-related biases that interfere with the correct use of a norm-free notion of actual cause.

There is an important tension between the first line of criticism and the other two. Suppose I am a proponent of the idiosyncratic (and potentially biased) view that the queen is in charge of watering the flowers and that the gardener is not supposed to water them. According to the adjust-the-model strategy, I am supposed to represent only those scenarios that I take to be serious possibilities. Thus, I will provide a model in which variable Q is the only cause of variable F. But this is a problem. Because now my idiosyncratic view does not only spoil my judgements of actual causation, but also the corresponding causal model!

The underlying point is this. Blanchard and Schaffer argue that the default/deviant distinction is unclear and reflects biases. But they also suggest to solve cases like the gardener/queen example by adjusting the models on the basis of considerations about

what scenarios are to be taken seriously. But what is a scenario that is to be taken seriously? Presumably this depends on ideas regarding normality that are similar to those affecting the defaults—otherwise it would be easy to generate counterexamples to the strategy. But this means that the constraint on models is no less unclear. In fact, exploiting constraints on $\mathcal{V}$ makes the problem even worse. For now the unclarities are not confined to the defaults but they infect the whole model. My argument in this paper is that there are situations where normality considerations should not affect the choice of variables in $\mathcal{V}$. If there are unclarities, then defaults are a better place for them.

2 The Problem of Disagreement

Consider a version of the gardener/queen case provided by Halpern and Hitchcock (2015):

> "while a homeowner is on a vacation, the weather is hot and dry, her next-door neighbour does not water her flowers, and the flowers die. Had the weather been different, or had her next-door neighbour watered the flowers, they would not have died" (414f).

Halpern and Hitchcock argue that since the flowers' death depends on both the weather and the neighbour's omission it seems like a counterfactual theory of causation cannot distinguish between these factors. Yet, according to some authors (e.g. Moore (2009)) the weather is a cause of the flowers' death but not the neighbour's omission to water them because omissions generally cannot be considered to be causes. Halpern and Hitchcock flag this as the "problem of isomorphism." But, according to them, there is "an even deeper problem. There is actually a range of different opinions in the literature about whether to count the neighbour's negligence as an actual cause of the flowers' death [...]. *Prima facie*, it does not seem that any theory of actual causation can respect all of these judgments without lapsing into inconsistency" (415). This is the Problem of Disagreement.

The Problem of Disagreement arises where the following two conditions hold. First, there are two (or more) agents that have conflicting judgements of actual causation with

regard to the same target system. For example, theorists like Moore argue that only the weather is an actual cause because they think that omissions cannot be actual causes. They disagree with theorists like Lewis (2000) who think that the neighbour's negligence is also an actual cause, because they think that omissions are genuine causes. Second, it has to be the case that the opposing agents agree on the underlying causal model. In the flower case Halpern and Hitchcock take this to be a model with a graph like the one in figure 1A, and a structural equation such that the flowers die if the weather is hot and the neighbour fails to water them: $D = H \wedge \neg W$.

What exactly do the agents disagree about in such cases? According to Halpern and Hitchcock, the disagreement concerns the actual cause of the outcome. But wouldn't this imply an implausible metaphysical view according to which actual causation is subjective? Halpern and Hitchcock admit that actual causation is a subjective and context-dependent notion that is to be distinguished from an underlying and objective notion of causal structure. Yet such a notion has an important function because it indicates targets of intervention that are particularly suited from the pragmatic perspective of the agent (Hitchcock and Knobe, 2009).

Let us see how Halpern and Hitchcock account for this case of disagreement. The idea is that the default/deviant distinction gives rise to a normality ordering over the worlds that can be represented by the model. Actual causes are those factors that fulfil the Halpern-Pearl (2005) definition of actual causation plus a normality criterion. The normality criterion requires that the possible world that is needed to show that the effect depends on the cause be at least as normal as the actual world. Halpern and Hitchcock argue that "[t]hose who maintain that omissions are never causes can be understood as having a normality ranking where absences or omissions are more typical than positive events" and Halpern and Hitchcock take this to reflect "a certain metaphysical view: there is a fundamental distinction between positive events and mere absences, and in the context of causal attribution, absences are always considered typical for candidate causes" (437f). This assumption of typicality gives rise to the judgement that the only actual cause of the flowers' death is the weather.

An advocate of the view that omissions are always causes can be understood as subscribing to an alternative normality ordering. Here the worlds in which the flowers do not die are equally normal and they are taken to be at least as normal as the world where the flowers die. Consequently, both the weather and the neighbour's negligence fulfil the normality criterion and qualify as actual causes of the flowers' death.

There are two problems with this reconstruction of the disagreement. First, it seems implausible that Moore would agree that absences are generally more normal than positive events. In fact, according to each of the many dimensions of normality, there seem to be clear counterexamples. Living humans more frequently breathe than not, functional smoke detectors remain silent (unless there is smoke), we are legally and morally required to help those whose lives are in danger. The kind of metaphysical point that Beebee and Moore make with regard to the causal status of omissions is independent of claims regarding the normality of omissions. Thus, it seems Halpern and Hitchcock have chosen an example where defaults do not do the explanatory work that they expect them to do.

Second, suppose for the sake of the argument that there is an agent who believes that absences are always considered typical and, thus, never can be causes. Moreover, suppose that the agent complies to the constraint that causal models should only represent scenarios that are to be taken seriously. According to the agent's beliefs, it is a very far-fetched possibility that omissions like the one of the neighbour are causes. Thus, the rules of appropriate modelling command that she leave out the variable representing the neighbour's negligence. But if this is the case, then this agent disagrees with the proponent of absences as causes already at the level of the standard causal models.

So, if we take the Problem of Disagreement to give rise to an argument for defaults, it seems like this argument faces the same difficulties as the argument from the Problem of Isomorphism. In particular, there is not really a problem in the first place if we choose what seem to be the most plausible representations of the agents' beliefs. The claim that there are agents who disagree about actual causes but agree on the underlying causal model seems to involve implausible empirical assumptions about the involved agents' sets of believes.

3 What is the Function of Extended Causal Models?

In the remainder of the paper I will argue that there is an alternative reading of the Problem of Disagreement that involves prescriptive claims about how disagreeing agents *should* use extended causal models. I will argue that the alternative reading shows that in some cases the default/deviant distinction is a useful extension.

What is the function of extended causal models? Halpern and Hitchcock "envision a kind of conceptual division of labour where the causal model [...] represents the objective patterns of dependence that could in principle be tested by intervening on the system, and [the normality ordering] represents the various normative and contextual factors that also influence judgments of actual causation" (2015, 435). So, it looks like causal reasoning involves considerations that are located at two distinct levels. First, there is the level of standard causal models. These represent the objective patterns of counterfactual dependence. Second, there is the level of judgements of actual causation. These judgements are influenced by the normality ordering which reflects normative and contextual considerations.

However, the conceptual division of labour does not seem to work as straightforwardly. First, objectivity on the level of standard causal models means that "once a suitable set of variables has been chosen, there is an objectively correct set of structural equations among those variables" (431f). Thus, the causal model itself is not objective. For the choice of the set of variables (and their possible values) is likely to be governed by criteria that are sensitive to normative and contextual factors as well (such as in Blanchard and Schaffer's treatment of the gardener/queen case). Second, even the judgements of actual causation need to have some objective core. Otherwise they could hardly help us "identify appropriate targets of corrective intervention" (432).

If causal models (plus information about the variables' actual values) and claims of actual causation are so similar, couldn't we just make do with one of them? No. Claims of actual causation are highly selective. And this has the advantage that they can guide agency very straightforwardly by indicating the best targets of intervention. Presumably, causal models are somewhat closer to the objective structure because

they allow representing larger chunks of it. They express complex counterfactual dependencies that are not captured by a simple claim of the form '$X = x$ is an actual cause of $Y = y$.' These larger chunks still depend upon norms, but do so to a lesser degree, for selection does not have to be constrained so narrowly.

In the following we shall see that the Problem of Disagreement helps to indicate one distinctive advantage of causal models, understood along these lines: they can help us provide a representation of disagreement of causes that is more conducive to resolving the disagreement than the bare claims of actual causation. Moreover, I shall argue that this function is sometimes (but not always) crucially facilitated by incorporating the default/deviant distinction.

4 An Example: Search and Rescue Missions

According to Frontex,[3] the European Border Control Agency, Non-governmental Search and Rescue missions (NGO SARs) are an actual cause of the increase of the number of deaths in the Central Mediterranean in the period from 2015 to 2016. On the other hand, it has been argued that NGO SARs are only one factor acting within a complex causal structure, and that it is erroneous to describe NGO SARs as *the* cause of the increase. I will look at a study performed by Forensic Oceanography[4] and show that the most natural way to understand their criticism of Frontex's claim is to see it as an attack on Frontex's assumptions about the causal model.

Let us begin with a closer look at the claims put forward in the Frontex report. The report describes an increase of the number of deaths of refugees and states that

> "it transpired that both border surveillance and SAR missions close to, or within, the 12-mile territorial waters of Libya have unintended consequences. Namely, they influence smugglers' planning and act as a pull factor that compounds the difficulties inherent in border control and saving lives at sea. Dangerous crossings on unseaworthy and overloaded vessels were organised

[3]The following is based on the risk analysis report for 2017 (FRONTEX (2017)).

[4]Forensic Oceanography is part of the Forensic Architecture agency located at Goldsmiths, University of London.

> with the main purpose of being detected by EUNAVFOR Med/Frontex and NGO vessels. Apparently, all parties involved in SAR operations in the Central Mediterranean unintentionally help criminals achieve their objectives at minimum cost, strengthen their business model by increasing the chances of success. Migrants and refugees – encouraged by the stories of those who had successfully made it in the past – attempt the dangerous crossing since they are aware of and rely on humanitarian assistance to reach the EU" (FRONTEX, 2017, 32).

Thus, the presence of SARs (both NGO and state-led operations) near the Libyan coastline is said to give a sense of security that encourages migrants and refugees to risk their lives. This has two effects. First, smugglers can offer crossings that are more risky. Second, there is an overall increase in attempted crossings.

The report also states that "[c]losely related issues are the safety of migrants and refugees and, most significantly, the increasing number of fatalities" (32). After reporting estimates of the fatalities in 2016 the report states that "[t]he increasing number of migrant deaths, despite the enhanced EUNAVFOR Med/Frontex surveillance and NGO rescue efforts, seems paradoxical at first glance" (33). But then the report relates the increase of fatalities to a change in the smugglers' tactics: "[t]he rising death toll mainly results from criminal activities aimed at making profit through the provision of smuggling services at any cost" (33).

It seems fair to assume that the above quoted passages can be summarized by the causal model displayed in figure 2A. In the model S represents the presence of SARs, C is a factor that represents the risk level of the individual crossing and the number of attempted crossings, and D represents the number of deaths. It is claimed that an increase in S leads to an increase in C, that an increase in C leads to an increase in D, and that an increase in S also leads to a direct decrease in D. The narrative does not allow a more detailed quantification of these functional relations. But there is a possible reading of the narrative according to which the increase of deaths via the route $\langle S, C, D\rangle$ is larger than the decrease via the route $\langle S, D\rangle$.

The Forensic Oceanography report (Heller and Pezzani, 2017) identifies the pull-

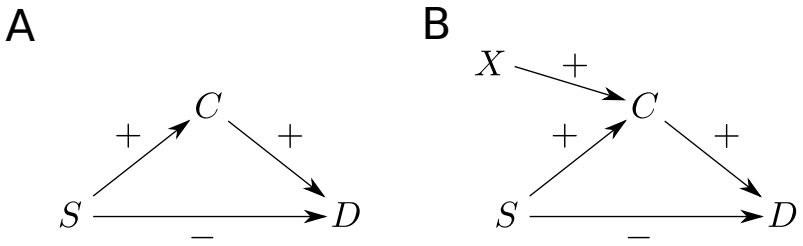


Figure 2: A simple model of the "pull-factor" narrative. A: Frontex. B: Forensic Oceanography.

factor claim as part of a toxic narrative within a "de-legitimisation and criminalisation campaign" directed at non-governmental search and rescue missions. The aim of the report is an empirical assessment of the claims put forward by Frontex. The report can be understood as challenging the structural relation between S and C as stated by Frontex and adding a variable X that feeds into C and that explains the increase in risk level and number of attempted crossings from 2015 to 2016. Relevant factors, among others, are the availability of seaworthy vessels, involvement of Libyan militia and Libyan Coast Guard.

The Forensic Oceanography report also describes non-governmental search and rescue missions as a continuation or replacement of preceding state-led search and rescue missions. In particular, the report claims that "[a]iming to deter migrants from crossing the Mediterranean, the EU and its member states pulled back from rescue at sea at the end of 2014, leading to record numbers of deaths. Non-governmental organisations (NGOs) were forced to deploy their own rescue missions in a desperate attempt to fill this gap and reduce casualties." That is, whereas the Frontex report suggests that there is a new kind of search and rescue missions that explains the increase, the Forensic Oceonography report describes the presence of search and rescue activity in the Mediterranean as a default condition.

5 The Role of Defaults

The disagreement between the Frontex report and the Forensic Oceanography report concerns the question whether the presence of SARs led to an increase in the number of

deaths. The underlying question is: why are refugees willing to risk their lives? The presence of SARs (and stories about how they guarantee safety on sea) is considered to be one factor. However, there are at least three further kinds of factors: (i) the situation in the home country, (ii) the hope for a better life in the EU, (iii) the absence of alternative pathways into the EU (legal pathways, or simply pathways that are not as risky).

Suppose each of these factors corresponds to a variable in a causal model such that a variable describing the willingness of refugees to risk their lives depends upon these variables. The disagreement about the causal role of SARs involves agents that have opposing views about which of these variables represent possible scenarios that are to be taken seriously—for functional, legal, and moral reasons. For example, there is disagreement about the moral and legal feasibility of cutting back life-saving missions on sea.

How *should* this disagreement be represented? One way would be to require that the involved agents agree on a set of variables $\mathcal{V}$ by including all variables that are at stake in the debate and represent their disagreement on the level of the default/deviant distinction. From a humanitarian perspective, for example, life-saving missions would be the moral and legal default state. By contrast, certain opposing agents might want to describe the absence of SARs as the default state. But both kinds of agents would be required to include a variable representing SARs.

Alternatively, one could require the views to be expressed by different standard causal models that reflect the individual views about what scenarios are to be taken seriously. This is what is suggested by the adjust-the-model strategy. The advantage is that such models do not incorporate the default/deviant distinction, which is considered unclear. The disadvantage, however, is that now the unclarity occurs in a disagreement about which scenarios are to be represented by the model in the first place.

The problem with this strategy is that it leaves unclear whether agents disagree for normative or for epistemic reasons. Suppose agent A_1 does not include a particular variable X in her standard causal model even though agent A_2 thinks that X is a cause of Y. Does agent A_1 mean to say that a change in X would merely amount to a scenario that is not to be taken seriously? Or does agent A_1 mean to imply that a change in

X would not make a difference to Y? Extended causal models fare better in this kind of context. They provide the formal resources that help the involved agents to point out where disagreement arises for normative reasons and where it arises for epistemic reasons. Agent A_1 would be required to include X into the model and clarify whether she takes Y to be independent of X or merely considers X to represent scenarios that from her particular point of view are highly abnormal.

This is particularly important in cases where it is likely that disagreement arises not only about norms but also about the underlying counterfactual dependencies. The core of Frontex's pull factor claim is the counterfactual dependency of C on S. This claim is difficult to assess directly. It involves non-trivial assumptions about the refugee's dispositions to risk their lives. It is also difficult to assess in an interventionist fashion. For performing testing interventions on the target system is unfeasible in practice. Instead Frontex supports the pull-factor claim by a comparison of the risk levels in 2015 and 2016 and relates this to an increase of the NGO SAR activity over this period. But this argument is valid only if all other potential causes for an increased risk level remain constant over this period. In Frontex's selective causal model it looks like this is the case. A more encompassing model such as the one provided by the Forensic Oceanography report, however, suggests that Frontex's claims are unwarranted. In order to warrant the pull-factor claim in the context of such a more encompassing model the Frontex report would have to show that the influence of these other factors is irrelevant.

6 Conclusion

The adjust-the-model strategy gives rise to a powerful objection to existing arguments for defaults that are based on the Problem of Isomorphism and the Problem of Disagreement. In this paper I have suggested a prescriptive reading of the Problem of Disagreement that provides a new argument for defaults that is not undermined by the adjust-the-model strategy. In cases of disagreement extended causal models should represent assumptions about the underlying causal structure that are shared by the involved agents, while the defaults should account for the normative disagreement. This helps keeping normative

disagreement apart from disagreement about the underlying counterfactual structure.

References

Blanchard, Thomas and Schaffer, Jonathan. Cause without default. In Helen Beebee, Huw Price, Christopher Hitchcock, editor, *Making a Difference*, pages 175–214. Oxford University Press, 2017.

FRONTEX. Risk analysis for 2017, 02 2017. URL `http://frontex.europa.eu/assets/Publications/Risk_Analysis/Annual_Risk_Analysis_2017.pdf`.

Hall, Ned. Structural equations and causation. *Philosophical Studies*, 132:109–136, 2007.

Halpern, Joseph Y. and Hitchcock, Christopher. Graded causation and defaults. *The British Journal of the Philosophy of Science*, 66:413–457, 2015.

Halpern, Joseph Y. and Pearl, Judea. Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

Heller, Charles and Pezzani, Lorenzo. Blaming the rescuers, 06 2017. URL `https://blamingtherescuers.org/`.

Hitchcock, Christopher and Knobe, Joshua. Cause and norm. *The Journal of Philosophy*, 106(11):587–612, 2009.

Lewis, David. Causation as influence. *The Journal of Philosophy*, 97(4):182–197, 2000.

Moore, Michael. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford University Press, 2009.

Pearl, Judea. *Causality. Models, Reasoning, and Inference*. Cambridge University Press, 2000.

van Fraassen, Bas C. *The Scientic Image*. Oxford University Press, 1980.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

Title: Edmond Goblot's (1858-1935) Selected Effects Theory of Function: A Reappraisal

Author: Justin Garson

Abstract: At the beginning of the twentieth century, the French philosopher of science Edmond Goblot wrote three prescient papers on function and teleology. He advanced the remarkable thesis that *functions are, as a matter of conceptual analysis, selected effects*. He also argued that "selection" must be understood broadly to include both evolutionary natural selection and intelligent design. Here, I do three things. First, I give an overview of Goblot's thought. Second, I identify his core thesis about function. Third, I argue that, despite its ingenuity, Goblot's expansive construal of "function" cannot be right. Still, Goblot deserves (long-overdue) credit for his work.

Keywords: Philosophy of biology; Edmond Goblot; biological function; selected effects

Address: Department of Philosophy, Hunter College and The Graduate Center, City University of New York, 695 Park Ave., New York, NY 10065.

Email: jgarson@hunter.cuny.edu

Acknowledgements: I wish to thank the participants of the workshop, "The Concept of Function in Biology: New Philosophical Perspectives," held at UQAM on October 4, 2019, including Brandon Conley, Antoine Dussault, Christophe Malaterre, and Parisa Moosavi. I particularly wish to thank the organizers, Antoine Dussault and Christophe Malaterre, for their extremely valuable feedback. I'm also grateful to Sarah Arnaud and Dan Dennett for their comments on an earlier draft.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

1. Introduction

According to the selected effects theory of function, a biological trait's *function* is, very roughly, whatever it was selected for by natural selection (or some comparable selection process). The function of the butterfly's eyespots is to deflect attack away from vital organs because that's what they were selected for. A chief virtue of the selected effects theory is that it makes sense of how function statements can work as teleological explanations – which is implicit in at least some strands of biological usage. If functions are selected effects, then when we attribute a function to a trait (say, deflection of attack to eyespots) we are quite literally offering an explanation for *why that trait exists*. No other account of function – perhaps with the exception of the "organizational theory" – even purports to make sense of this feature of biological usage.

A consensus among philosophers of biology is that the selected effects theory was first formulated independently by Neander (1983) and Millikan (1984), though perhaps earlier work such as Wright (1973), Wimsatt (1972) and Ruse (1971), gestured in that direction. One goal of this paper is to challenge that consensus. The forgotten French philosopher of science, Edmond Goblot (1858-1935), should be credited with formulating the theory, or at least an early incarnation of it. In a series of papers, Goblot (1899; 1900; 1903) argued, quite rigorously and explicitly, first, that function statements are teleological explanations, and second, that function statements can be teleological explanations only if *functions are selected effects*. My goal is not in any way to undermine the originality and insight of Neander, Millikan and their followers. It is rather to ensure that Goblot receives long-overdue credit for his prescient discovery.

But this paper does not simply have the goal of insisting that Goblot receive some intellectual credit. That fact alone would be worthy of an extended footnote in a philosophy of biology textbook, and not a whole paper. Goblot, however, did much more than that. He articulated a very distinctive (even by today's standards) version of the selected effects theory. For Goblot, "selection" was much more inclusive than evolutionary natural selection. It was even more inclusive than the abstract notion of "differential reproduction," or "differential retention," as some would have it. For Goblot, "selection" refers to a very general process wherein *one possibility is realized, to the exclusion of another, by virtue of an apparent advantage*. For Goblot, evolutionary natural selection, and intelligent design, are two subtypes of this abstractly-specified process.

Moreover, Goblot seemed to think that *this* claim – that functions are selected effects (when "selection" is broadly construed) – is *a conceptual analysis of both lay and scientific use of "function."* If Goblot were right, that would be game-changing even by today's standards. For it would imply that the selected effects theory, properly grasped, embraces both biological and artifact functions, scientific and lay usage, modern and ancient usage. This expansive construal of the selected effects theory deserves serious consideration.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

To be fair, some philosophers of biology have flirted with expanding the selected effects theory to be more inclusive, that is, to allow processes other than evolutionary natural selection to produce new functions. These thinkers include Millikan (1984) herself, Papineau (1984), Godfrey-Smith (1992), Griffiths (1993), Kitcher (1993), and Garson (2011). Dennett (1969), Wimsatt (1972), and Wright (1973) also gestured toward the possibility of such an all-encompassing theory. Goblot, however, is unique in that he joined two ideas that nobody else joined: First, he attempted to specify, rigorously and precisely, the nature of this general process *of which* natural selection and intelligent design are subtypes. Second, he posited that this fact, that functions are selected effects – when selection is understood in this expansive way – is part of a correct conceptual analysis of "function." This is a new thing.

Unfortunately, it seems to me that, while there's something right about Goblot's expansive way of thinking about functions, his particular construal of function cannot be right. That's because there's no single *kind* of process in the world of which natural selection and intelligent design are subtypes. The illusion that there is a single kind of process in nature arises from a hidden equivocation in the very idea of "selection for an advantage." As I'll show, one sense of the phrase points to human choice; the other to evolutionary natural selection; these – as Darwin himself recognized – cannot be fused in any non-metaphorical way. Though Goblot's attempt fails, it's a quite *noble* sort of failure, one that still demands a serious philosophical reckoning.

2. Goblot's Basic Account of Function and Teleology

Goblot wrote two major papers on the topic of function and teleology, "Fonction et Finalité" of 1899 and "La Finalité Sans Intelligence" of 1900.[1] Crucially, Goblot intended the two papers to be read as a continuous whole. This can be seen from the fact that the purpose of the first paper is to raise a general problem about biological functions, and the purpose of the second paper is to solve that problem. In fact, the first paper actually ends with the parenthetical remark "*A suivre*" – "to be continued." This is important for us, because it helps us to see that the two papers are intended to be read as one long meditation on functions.

Though the two papers are meant to be read as one, each pursues a distinct question and offers a distinct thesis. The first paper, "Fonction et Finalité," argues that *function statements are teleological explanations.* When we say, for example, "the function of the eyespots on butterfly wings is to deter attacks away from vital organs," we are, in ordinary biological discourse, trying to explain *why butterfly wings have eyespots*. The second paper, "La Finalité Sans Intelligence," argues that teleological explanations are grounded (in a way to be determined) by evolutionary natural selection. Hence, on the surface, his position seems nearly identical to that which Larry Wright developed in

[1] A third paper, his "La Finalité en Biologie" of 1903, is a commentary on other works and will not be discussed here.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

1973: function statements are teleological explanations (a trait has a function if the trait "is there because" it serves the function), and one way that a trait can have a function is if it was shaped by natural selection for the effect in question.

As we will see, however, Goblot goes much further by arguing that functions *are just* selected effects, where selection must be construed broadly enough to include both evolutionary natural selection and intelligent design.[2] Before diving deeply into Goblot's analysis, I'll turn to the text to draw out Goblot's own presentation of these two theses. Any further analysis we conduct must be based squarely on Goblot's own words.

His first paper argues that the function of a trait is not just any useful effect it happens to have. Rather, a trait's function is the effect that the trait (in some sense) was *made for*. It's an effect that plays into an explanation of the trait itself. He begins his analysis by pointing out that functions, in the ordinary biological sense of the term, are peculiar and worthy of serious philosophical reflection:

> Of the properties of cells, tissues, and organs, some are, and others are not, *functions*. Sometimes scientists intentionally use this word *function*; sometimes on the contrary they take care to avoid it; the definition is difficult, but the use is not at all arbitrary (1899, 495).[3]

He then argues that, in ordinary biology, we only call something a "function" when we think that the effect in question is somehow part of an explanation for the trait's existence:

> The blood cell fixes atmospheric oxygen; it also fixes carbon monoxide and nitrogen dioxide…Of these three chemical properties, only the first one is a function; and the only reason that one calls it that, is that the cell is made to draw, in its passage in the lungs, atmospheric oxygen…If the cell also fixes other gases, these properties are not functions, for it *is not made for that* (1899, 497-8; emphasis in original).

The problem, of course, is that it is very difficult to see how an *effect* of a trait can be part of the explanation for that very trait, unless we are invoking some sort of supernatural principle, such as divine intervention or a mysterious vital force:

---

[2] One might think that this is precisely what Wright (1973) was saying, particularly because of his suggestive comments on pages 162-4 about the similarity of the concept of *selection* in natural selection and in intelligent design. One would be mistaken, for reasons to be discussed in Section 4. *Wright did not think that functions were selected effects*, regardless of whether "selection" is construed narrowly or broadly. This was, of course, a major point of Neander's (1983) and Millikan's (1984) critiques. See Garson (2016, Chp. 3) for more on the relevant historical background.

[3] All translations from the French are my own.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

> Certain physiologists seem to have a sort of distrust for this idea of finality, which, despite them, can be found in all parts of their science. They dare not look at it directly; finality seems unknowable; for them, it is an anti-scientific, and almost mystical, idea (1899, 499).

Nonetheless, teleology is such a critical part of physiology itself that if we eliminate teleology, we eliminate physiology, too:

> Does there exist, in the facts, a teleological order? Put differently, is physiology possible?...The existence of a teleological order is the postulate of the science of life…The physiologist must therefore assume the reality of a teleological order, as the physicist assumes the reality of a constant and necessary order (1899, 504-5).

That is the puzzle that his paper ends with: teleology seems both impossible and necessary.

The purpose of his next paper, "La Finalité Sans Intelligence," is, as the title indicates, to point the way to a solution. If, in the past, a trait was shaped by evolutionary natural selection for a certain effect, *then that trait exists now precisely because of that effect*. If the flower's nectar glands were selected for attracting insects, then we can rightfully say, *now*, that the nectar glands exist (that is, one reason flowers have nectar glands) *because* they attract insects. When selection is present, a trait's effect can be cited as part of an explanation for its existence, without appealing to theism or vitalism.

> But if it happens that an individual character is an *advantage*, natural selection will make of it a species character, and that because it is an advantage. Hence again there is finality, but finality without intelligence…It is easy to see that these examples [e.g., "the function of nectar glands in flowers is to attract insects"] answer to the definition of finality, for the consequent is the *raison d'être* of the antecedents. Cross-fertilization exists because it causes greater fecundity; nectar glands, large or brilliant corollas, perfumes exist because they have the effect of attracting insects...It would not be exact to say that the effect is here the cause of its cause, but it is true to say that it is the reason for it; the existence of the cause is explained by the effects that it produces (1900, 402-3).

And later:

> [After selection,] the final term [that is, the trait with the function which is now a "fixed" species character] no longer has an accidental character, since it is this very advantage, which has become a species character. Utility is the origin of finality; utility characterizes the initial term, it *serves* a certain end, but it is not made for this end; finality characterizes the final term; it is well made for this usage, since it is because of its utility that it became fixed as a species character. (1900, 404).

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

In sum, Goblot holds that function statements are teleological explanations, and that natural selection can vindicate such explanations, since natural selection shows how a trait's effect can play a role in an explanation of that trait's existence. As we will see, however, Goblot thinks that the relation between function and selection is, in fact, even more intimate than this.

3. Rethinking Teleology as Selection

What is really innovative about Goblot's thought, even by today's standards, is what I take to be his core thesis about biological functions, one that is never stated explicitly but implied throughout the text.[4] I will try to articulate the view as follows: *As a matter of conceptual analysis, an object has a function when it is the result of an abstract kind of selection process. In this selection process, one possibility is realized to the exclusion of another on account of something like "the appearance of an advantage* [l'apparition d'un avantage]*." The function of the object is just this advantage. Evolutionary natural selection, and intelligent design, are two different subtypes of this abstract process.*

My main textual evidence that this was, in fact, Goblot's view of function, stems from the extraordinary closing section of "La Finalité Sans Intelligence." There, he states that *all* teleology, intelligent or not, somehow involves *a selection between possibilities and the preferential realization of one over another*:

> All finality, intelligent or not, is a *choice* between possibilities…Natural selection is the effective *trial* of all of the possibilities. The one which is the best wins only by proof of its superiority. Intelligent finality is more rapid and economical, since the possibilities are judged before being tried; or rather, the trials are made ideally instead of being carried out. It is also therefore a sort of selection, which operates between ideas. The God of Leibniz conceives in thought all of the possible worlds; he compares them, judges them, and realizes the best...There is therefore, in the divine understanding, *competition* between the possible worlds and *selection* of the best. Things are no different in our own deliberations. There is a competition between the diverse choices we can make, and selection of that which is or which seems to us the best. The initial term is always the appearance of an advantage; the final term the realization of this advantage. The analogy is therefore complete between intelligent and unintelligent finality; only intelligence abridges the path and diminishes the effort. Finality, therefore, is not at all the characteristic mark and like a seal of intelligence imprinted on its works. Intelligent finality is a specific mode of finality in general. (405-6)

[4] Bonsack (1976) is the only paper that I've encountered that critically engages with Goblot's teleology. His main complaint is that Goblot defines *finalité* differently in different places, and that he introduces inappropriate value notions. I agree with thrust of his critique, but I find a unified notion of biological function underlying Goblot's presentation. (I thank Antoine Dussault for drawing Bonsack's paper to my attention.)

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

In this passage, Goblot leads us through three major areas where teleological statements loom large, and shows us that, in each of the three domains, teleology exhibits the same fundamental pattern. First, he asks us to consider intelligent design in the creationist worldview. Suppose we are willing to agree that some feature of the world is designed by God for certain end. We can then ask ourselves: what exactly is God doing when God "designs" something? It consists in none other than this: God somehow *surveys* a vast array of possibilities, and *chooses* to realize one possibility over another because of an apparent advantage.

In the lengthy passage cited above, Goblot is quick to point out that the same pattern is exemplified in *human* decision-making. When a person designs something, something takes place in her mind that is like a competition between imagined possibilities, and one possibility is ultimately realized, over another, because of an apparent advantage.[5]

Finally, and most importantly, Goblot sees evolutionary natural selection as conforming to this basic pattern. Natural selection, he thinks, involves a *competition between possibilities*, wherein one possibility is realized, over another, because of the appearance of an advantage. At this juncture, one might suspect that Goblot is playing a semantic game with us, or that he is abusing the natural contours of ordinary language. Surely, natural selection isn't a competition between *possibilities*! To the extent that natural selection is a "competition," it's a competition between *actual* organisms (cells, groups) and not merely *possible* ones.

Though natural selection must be seen as a competition between *actual*, rather than *possible*, beings, for Goblot, natural selection is also, and at the same time, a competition between *possible species characters*. When a new variant arises in a population through a genetic mutation – say, the first butterfly with eyespots on its wings – that variant represents a *possible species character*. It is not yet an actual species character; it must compete with other variants to earn that title. One thing that natural selection does is that it takes a possible species character and transforms it into an actual species character because of an advantage it possesses:

> But if it happens that an individual character is an *advantage*, natural selection will make of it a species character, and that because it is an advantage. Hence again there is finality, but finality without intelligence…(1900, 402)

[5] Christophe Malaterre has pointed out to me that the French text admits of a different interpretation, where "our own deliberations [nos propres délibérations]" refers to an *inter*personal, rather than *intra*personal, decision-making process. For example, we can speak of a committee "deliberating over" various social policies. This interpretation would still imply that the kind of function a social policy has is the same *kind* of thing as the kind of function that a biological organ has.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

It is because of this somewhat unconventional perspective on natural selection that Goblot can see it as conforming to the basic pattern of teleology in other domains.

4. Convergence and Divergence

As I noted earlier, Goblot is not the only person to suggest a deep similarity between natural selection and intelligent design. Many philosophers have hinted at a deep connection, even identity, between the two sorts of things. It is impossible to do justice here, in a rather short paper, to the rich similarities and differences between these theorists. Here it will have to suffice to say this: nobody, with the exception of Goblot, has ever held this convergence of ideas:

(1) Functions are selected effects.
(2) "Selection" in (1) must be understood very generally to encompass natural selection and intelligent design.
(3) (1), understood in terms of (2), is a conceptual analysis of both ordinary and scientific language.

An all-too-brief perusal of the literature will show exactly how and where Goblot departs from others. To begin with, Wright (1973) didn't accept (1), at least not as a conceptual analysis. He thought that, as a matter of conceptual analysis, a function of a trait is just an effect that explains the trait's existence. He does discuss the similarity between natural selection and intelligent design, and even the idea that they both exemplify, in a very abstract way, a kind of "selection process" (see pp. 163-4), but he never *identifies* function, as a matter of conceptual analysis, with this abstract "selection process." He identifies it with what he calls a "consequence-etiology."

Wimsatt (1972, 13) seemed to accept that, empirically speaking, functions probably always involve selection, where "selection" is understood broadly to encompass both natural selection and intelligent design: "the operation of selection processes is not only *not* special to biology, but appears to be at the core of teleology and purposeful activity wherever they occur." But he adamantly rejected that this should be understood as a conceptual analysis.

Dennett (1969), too, describes a deep analogy between natural selection and learning, and even suggests that selection is at the root of teleology itself (64), though he does not develop this insight into a theory of "function" *per se*. In fact, Dennett has pursued this analogy throughout much of his work, particularly in his classification of "Darwinian," "Skinnerian," and "Popperian" creatures (1995), each of which involves the operation of different sorts of selection processes.

Millikan (1984) defined functions in terms of a general process involving the differential reproduction of one type of entity over another. Her view of "reproduction" is expansive enough to include trial-and-error learning and learning by imitation (p. 28). But in her view, differential reproduction does *not* include the process wherein a person creates an

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

artifact for the first time. (The first hammer did not have a history of differential reproduction on account of its past success.) Goblot's view is, therefore, more inclusive than hers. Similarly, Papineau (1984, 557-8) states that mental states can undergo natural selection within the lifetime of an individual, and thereby acquire (selected effects) functions – but not, presumably, the artifacts that are produced by said mental states.

Kitcher (1993), to whom I'll return in the next section, says that function is "design," as a matter of conceptual analysis, and that natural selection and intelligent design are two subtypes of this "design." Unlike Goblot, however, he does not articulate what "design" is supposed to be *such that* it encompasses both. That is, he never articulates what this non-metaphorical form of "design" is supposed to amount to, other than alluding to Darwin's view that natural selection can be seen as a kind of "design without a designer."

Neander (1991) holds that functions are selected effects, but only in the sense of Darwinian natural selection. Hence, her theory of function is only supposed to apply to the biological sort of function, and it is only intended as a conceptual analysis of *modern biological usage*. She notes, in passing, the possibility of a theory like Goblot's, but chooses not to develop it in any detail (p. 175).

Griffiths (1993) sketches a theory of artifact function that rests on the idea that artifacts come from a kind of "competition" of ideas. But he says explicitly that natural selection and artifact design are quite different things and that there is no single concept of function that applies to both (p. 421).

5. A Critique

If Goblot were right, that would be a game-changer for contemporary philosophical discussion of function and teleology. That is because it would give us a version of the selected effects theory that effortlessly captures teleology in every domain in which it arises, both natural and conscious, human and divine. It would also, as a conceptual analysis, unify both modern and ancient usage, as well as scientific and lay usage. For Goblot, by "function," *everybody has always meant selected effect*.

Unfortunately, Goblot's expansive analysis of function simply does not work. The reason is that *there is no single kind of process in the world, loosely called "selection for an advantage," of which both natural selection and intelligent design are two subtypes*. There is only a strained analogy. Goblot's view relies, ultimately, on an unacceptable anthropomorphism.

The crux of the matter is this: in the standard selected effects theory, an object's effect becomes that object's function by virtue of the fact that that sort of object has an *actual, historical, track record* of producing that effect. Having an *actual, historical track record* of producing a given effect is necessary for having a function, in the ordinary selected effects sense. Artifacts, however, are not subject to this constraint. As far as artifacts go, it is possible for the effect of some artifact to be its function even if that artifact has no

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

actual, historical track record of producing that effect. Neander (1991, 174-5) makes precisely this point in enumerating the differences between natural and artifact functions. The very first twist corkscrew that was ever invented, back in 1795 by the Reverend Samuel Henshall, no doubt had the function of opening wine bottles, even though it had no actual historical track record of doing so. True, there may have been a kind of "virtual selection process" involved in the production of the first spiral corkscrew, a selection process that took place in the Reverend's mind. *But the physical corkscrew, that is, that physical type of thing, was not selected because it actually ever opened a wine bottle, since it had never done so*. It was selected because someone (namely, Henshall), *thought*, or *surmised*, or *believed*, or *reckoned*, or *figured*, that it would have that advantage. But anticipated advantages are not real advantages, any more than imaginary ponies are real ponies. To say that all function involves something like "selection for an advantage" obliterates that distinction.

Let me put the point somewhat differently: Goblot's argument involves a fallacy of equivocation. In the fallacy of equivocation, two or more premises only seem to support a conclusion because of a critical ambiguity in a word or phrase that appears in the premises. Goblot, I maintain, is guilty of such an equivocation. We can reconstruct his argument as follows: *natural functions involve selection for an advantage; artifact functions involve selection for an advantage; so, natural and artifact functions both involve selection for an advantage*. The ambiguity is this: in the first premise, the "advantages" in question are real, actual advantages; in the second, the "advantages" in question are merely imagined or hoped for. But imagined advantages are not real advantages – any more than imagined ponies are real ponies.

The problem is analogous to the problem in Kitcher's (1993) theory of function. Kitcher attempts to define "function" simply and solely in terms of "design." He then says that human invention, and Darwinian natural selection, are two subtypes of this "design." The problem is that the apparent unity of the concept of function is purchased at the cost of an equivocation: there is no single kind of thing called "design," of which human choice and natural selection are subtypes. The latter is "design" in name only; it is a clever anthropomorphism to speak of natural selection as a form of design, but this analogy cannot bear any real theoretical weight.[6]

6. Conclusion

If Goblot is wrong, then what is the right way to think about function and selection? First, if we are to maintain that functions are selected effects, we should continue to understand "selection" in a *relatively* narrow sense which requires (not as a sufficient condition, but as a necessary one) something like an actual history of differential reproduction, or differential retention, on account of the effect in question. It is not enough that some agent *hopes* or *anticipates* or *surmises* that the object will have the relevant effect. In

[6] That said, there is much to appreciate in Kitcher's view, in particular the distinction between selection having a "direct" versus an "indirect" role in a trait's function.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

contrast, functions in the realm of *artifacts* come about because the artifact bears the right kind of relationship to an agent's mental states – though the precise nature of that relationship remains highly contentious. But I think we should give up the search for a unified theory of biological functions and artifact functions. Too many smart people have tried and failed for that to be a fruitful endeavor.

Now, it may very well be true that, as a rule, when someone produces an artifact, that event of production is preceded by something like a *virtual selection process* in the agent's mind. Dennett (1995) calls us "Popperian creatures" for our ability to carry out a hypothetical trial-and-error in our minds before implementing our schemes in the real world. But this, in my view, is incidental to an artifact's having a function. It is not by virtue of the fact that a selection process takes place "in the designer's mind" that the artifact acquires a function. As Wimsatt (1972, 15-16) argued some time ago, if God is real, and if God had a creative hand in designing the universe or some of the things in it, he wouldn't have had to go through anything like a virtual form of trial-and-error in order for his creations to have functions. He would have just known what to do.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

References

Bonsack, F., et al. 1976. *Fonction et Finalité: Symposium Ecrit*. Bienne, Switzerland: Association F. Gonseth (Institut de la Méthode).

Dennett, D. 1969. *Content and Consciousness*. London: Routledge.

Dennett, D. 1995. *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Garson, J. 2011. Selected effects and causal role functions in the brain: the case for an etiological approach to neuroscience. *Biology and Philosophy* 26:547-565.

Garson, J. 2016. *A Critical Overview of Biological Functions*. Dordrecht: Springer.

Goblot, E. 1899. Fonction et finalité. *Revue Philosophique de la France et de l'Étranger* 47: 495-505.

Goblot, E. 1900. La finalité sans intelligence. *Revue de Métaphysique et de Morale* 8: 393-406.

Goblot, E. 1903. La finalité en biologie. *Revue Philosophique de la France et de l'Étranger* 56: 366-379.

Godfrey-Smith, P. (1992). Indication and adaptation. Synthese 92: 283–312.

Griffiths, P. E. (1993). Functional analysis and proper function. *British Journal for the Philosophy of Science* 44: 409–422.

Kitcher, P. 1993. Function and design. *Midwest Studies in Philosophy* 18: 379-397.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Neander, K. 1983. *Abnormal Psychobiology*. Dissertation, La Trobe.

Neander, K. 1991. Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.

Papineau, D. 1984. Representation and explanation. *Philosophy of Science* 51: 550-72.

Ruse, M. E. 1971. Functional statements in biology. *Philosophy of Science* 38: 87-95.

Wright, L. 1973. Functions. *Philosophical Review* 82: 139-168.

Copyright Philosophy of Science 2020
Preprint (not copyedited or formatted)
Please use DOI when citing or quoting

Wimsatt, W. C. (1972). Teleology and the logical structure of function statements. *Studies in the History and Philosophy of Science* 3: 1–80.

Causal and Non-Causal Explanations of Artificial Intelligence

Christopher Grimsley

4,818 words

Abstract

Deep neural networks (DNNs), a particularly effective type of artificial intelligence, currently lack a scientific explanation. The philosophy of science is uniquely equipped to handle this problem. Computer science has attempted, unsuccessfully, to explain DNNs. I review these contributions, then identify shortcomings in their approaches. The complexity of DNNs prohibits the articulation of relevant causal relationships between their parts, and as a result causal explanations fail. I show that many non-causal accounts, though more promising, also fail to explain AI. This highlights a problem with existing accounts of scientific explanation rather than with AI or DNNs.

1 The Need for Explainable Artificial Intelligence

The use of artificial intelligence (AI) has expanded considerably in the past decade. AI is increasingly being used to make high-stakes decisions, often under questionable circumstances that indicate the presence of racial or gender bias, including granting or denying loan applications (Fuster et al. 2018), deciding which prisoners are eligible for parole (Khademi and Honavar 2019), and diagnosing mental health disorders (Bennett et al. 2019). If AI is used to make these decisions — especially if these decisions appear to have reinforced biases present elsewhere in society — understanding how the algorithm made the decision is essential. Absent explanation, arbitrary or biased decisions may go unchecked. Computer scientists have recognized this problem and have begun developing explainable AI (XAI), but many of their strategies haphazardly employ a mix of causal, psychological, and counterfactual strategies that fail to generate adequate explanations. It is impossible to explain AI without first explaining explanation. The philosophy of science is uniquely positioned to take on this problem and offer solutions by examining the meaning of scientific explanation and developing an account of explanation which adequately explains AI.

An explainable algorithm is one for which a true, satisfactory explanation exists. An interpretable algorithm is one for which a complete account of the relationships between the steps in the algorithm exists. In many cases, AI decision and classification algorithms are neither explainable nor interpretable. Many of the AI algorithms used in these cases are deep neural networks (DNNs), a type of algorithm whose complexity defies explanation in a particularly striking manner. Because explanation through merely technological means is lagging behind the complexity of the networks that are in

need of an explanation, it is reasonable to conclude that the solution to this problem cannot be technological. If this is the case, a potential solution can be found in the ways in which explanation is conceptualized within the context of AI. In order to solve the explainability problem, it is first necessary to articulate an appropriate model of explanation which can be effectively applied in this context.

I argue that recent attempts by computer scientists to develop XAI fail because they do not employ a theoretically-grounded concept of explanation. Further, I show that it is necessary to employ non-causal accounts of explanation in order to solve the problem of explainability in AI. I begin with a brief overview of the aspects of AI that are relevant to my argument. Then I discuss two existing methods for developing XAI: one causal, and one non-causal. I demonstrate why each approach fails to generate a satisfactory explanation, then I propose alternative non-causal possibilities and explore the viability of each. I conclude that existing approaches to both causal and non-causal explanation fail to fit the needs of XAI, though of the two approaches, non-causal accounts hold greater promise.

1.1 Deep Neural Networks

'Machine learning,'[1] an increasingly common form of AI, is a broad term that describes programs that can work with unexpected input data without being explicitly programmed to do so. One of the more common contemporary approaches to machine learning is the neural network. Neural networks attempt to replicate the behavior of biological brains by linking input and output together via various intermediary nodes in

[1]for a more comprehensive overview, see Buckner (2019).

a network. Each node is called a 'neuron', hence 'neural network'. Neural networks contain multiple layers including an input layer, an output layer, and one or more 'hidden layers' between the input and output. Each layer is made up of a group of neurons. Neural networks with more than three hidden layers are called deep neural networks (DNNs). DNNs produce a complex, often non-interpretable model that is used in decision or classification tasks. In what is called 'supervised learning,' a 'trained model' is created by providing labeled datasets to the DNN, which iterates over the labeled data and builds a model capable of making the correct decision or classification given novel data. In other words, the deep neural model is built with the deep neural network. DNNs and the models they produce are both in need of explanation.

2 The Current Landscape: Two Case Studies

Computer scientists have made use of two contrasting strategies in order to develop XAI. Most researchers attempting to build explainable DNNs appear to prefer causal forms of explanation,[2] however some have attempted to develop non-causally explainable DNNs. I present instances of each approach and discuss their relationships to the explanation literature in the philosophy of science.

2.1 Case Study One: "Rationalizations"

One approach to XAI is to develop algorithms that produce patterns of explananda that imitate human reasoning. This is analogous to chatbots that imitate human texting

[2]See for example Yang et al. (2016), Jain and Wallace (2019), Khademi and Honavarand (2019), and Sharma, Henderson, and Ghosh (2020)

patterns. For instance, Harrison et al. (2017) uses two AIs. The first plays the classic video game Frogger, and the second explains the actions of the first by translating internal game state data to natural-language approximations of human-supplied explanations. In order to accomplish this, the research team recorded human subjects playing Frogger, then periodically paused the game and asked the subjects to verbally explain an action that they recently took. The human responses were used as training data for the "explainer" DNN.

Importantly, the explainer DNN was not generating veridical statements about the internal state of the game-playing DNN, but was generating a unique natural-language statement based on data gathered from human players when in similar in-game situations. This approach generates psychologically satisfying explanations of AI behavior. Because the generated explanations are only meant to approximate human-supplied explanations of similar situations, a tradeoff is made between accurately reporting internal DNN states and psychologically satisfying explanations. The authors accept this tradeoff in order to obtain quickly-generated and human-like explanations. The authors write that "rationalization is fast, sacrificing absolute accuracy for real-time response" (Harrison et al. 2017, 1).

The explainer DNN does not supply a veridical explanation of the decision making process used by the game-player DNN. Instead it produces statements that approximate human-generated explanations when faced with similar in-game circumstances. Another much deeper problem with this model is that, since the explanation of one DNN is itself generated by a different, independent DNN, there is now a need for an explanation of the explanation. If one black-box system is explained by appealing to a second black-box system, nothing has actually been explained. The number of phenomena in need of

explanation has actually increased.

If humans depend on the use of AI for a critical task, it is important that a sense of trust in that AI is maintained. One goal of the research of Harrison et al. (2017) is to provide explanations that reassure human operators of AI that the AI had a good reason for doing an action that may appear to a human to be questionable. In some cases this may mean that the AI only needs to be able to communicate that a good reason for a particular action exists, i.e. to articulate a how-possibly explanation, rather than communicating the right reason for the action, i.e. a how-actually explanation.

Rationalizations are an attempt to deal with the problems associated with the lack of XAI without actually solving them. The authors endorse the view that, when it comes to AI, we must choose between fast, intuitive, human-understandable explanations, and technically correct explanations. Rationalizations do not attempt to provide explanations, but instead provide fictional statements that sound like plausible explanations.

2.2 Why Rationalizations are not Explanations

Rationalizations represent only one attempt to build non-causal XAI, but this attempt leaves much to be desired from the standpoint of scientific explanation. Rationalizations are explicitly non-veridical. Fictionalizations often serve a role in scientific explanation. Many, including Potochnik (2017) and Rice (2018), have argued that fictionalizations can play a key role in understanding. Rationalizations differ from fictionalizations in other models. If the understanding that an explanation helps to foster is not in any sense an understanding of a true state of affairs, then the purported explanation has not

contributed to epistemic success, and is not actually explanatory. Rationalizations do not make use of strategic inaccuracies in order to help individuals to come to recognize a greater truth about the explanandum, rather rationalizations serve to further conceal the truth behind natural language statements meant to have the appearance of an adequate explanation with none of its substance. While there may be practical reasons why AI developers would find it appropriate to make use of rationalizations rather than genuine explanations, this does not imply that rationalizations have any value as scientific explanations. Rationalizations are an attempt to articulate "how possibly" explanations rather than "how actually" explanations. In the case of explanations of high-stakes automated decisions, "how actually" should be the standard. Rationalizations are not explanations.

2.3 Case Study Two: Attention Layers in Neural Networks

Attention mechanisms, introduced by Bahdanau et al. (2015), allow the training of a DNN in such a way as to focus the network's attention on specific input elements. Attention mechanisms can be incorporated into neural networks as another layer of the network as shown in figure 1. The weights of the attention layer are thought to correlate to measures of feature importance in the input: the input has some features that are more important than others, and if the attention layer is able to identify which features of the input are most important, this is thought to generate explanantia by discriminating between relevant and irrelevant inputs. Allowing the DNN to focus on the more important parts of the input could increase the accuracy of the output. In the case of attention as explanation, the explanandum is the output of the DNN, and the

explanans involves an appeal to the attention layer, which points to specific input elements. In many cases, it appeared as if the attention layer was explanatory because it indicated which parts of the input were most important in the creation of the output. For those evaluating these systems for explanatory value, this appears to be a plausible explanation, though as I will discuss, there are good reasons for doubting that this is true.

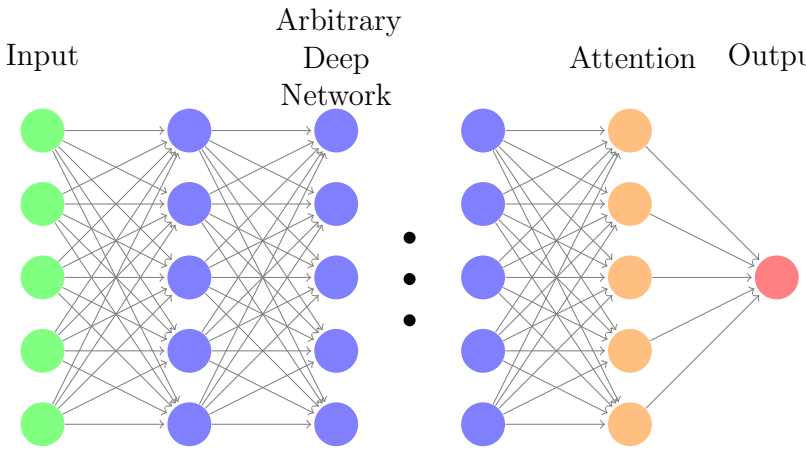


Figure 1: Researchers often use attention weights (shown in orange) to generate explanations. Jain & Wallace scramble attention weights and show that output remains stable; a similar result is obtained by Serrano & Smith omitting highly-weighted attention nodes entirely.

2.3.1 Critical Responses from Computer Science

Jain and Wallace (2019) argue that the output of the attention layer cannot serve as an explanation of the underlying DNN because it is possible to intentionally interefere with the way the weights of the attention layer are set (called "adversarial weighting") in such a way that the underlying DNN produces the same output as it did under non-adversarial weighting while the adversarial attention layer indicates the importance

of entirely different - and obviously unimportant - elements of the input data. An example discussed by Jain and Wallace is the use of a DNN to gauge whether a movie review is positive or negative. The DNN outputs a number between 0 and 1 with 0 being very negative and 1 being very positive. The attention layer indicates which words in the movie review (the input) are supposedly more important in determining this output. Under the non-adversarial case, a word like "waste" would be indicated as important, whereas under the adversarial weighting, a word like "was" would be indicated as important. In both the adversarial and non-adversarial cases, the network produced an identical score for the review.

While the attention weights were set adversarially, they still represent a configuration that could have occurred during the non-adversarial training of the network. In developing a neural model under normal conditions, the production of either of the models (adversarial or non-adversarial) are equally possible. If one expects that the attention layer can serve as an explantion of the overall model, it must be the result of the ability of the attention layer to identify the most important features of the input data, but if selectively randomized attention weightings can produce the same model output as the actual attention weights, it is difficult to see in what sense the attention layer could possibly generate an explanation. Jain and Wallace (2019) conclude that it cannot. Their paper is appropriately titled "Attention is not Explanation."

Serrano and Smith (2019) make a similar argument, agreeing that attention is not explanation. Instead of assigning randomized weights to the attention nodes, Serrano and Smith selectively deleted many of the highest weighted - that is the supposedly most important - attention nodes. Under these conditions the model still produced the same output. The experiment demonstrates that if adversarial attention weightings using data

that should adversely affect the neural model's accuracy has no such effect, the ability of the attention layer to discriminate between important and unimportant inputs is called into question, and so must be any explanations that are derived from attention.

Both of these papers relied on counterfactual analyses of the attention layer in order to come to thier conclusions: if the attention weights had been different in such and such a way, the attention layer would have identified a different set of input features, while the model's output would have remained unchanged. Implicitly, both are appealing to an interventionist account of explanation. They are attempting to determine the pattern of counterfactual dependence among the variables in the DNN. As I show below, due to the complexity and lack of interpretability of the systems this analysis is being applied to, the use of the interventionist account here is inappropriate, and is not likely to lead to the development of XAI.

2.4 Why Attention is not Explanation

Alisa Bokulich (2018) defines 'causal imperialism' as the view that "all scientific explanations are causal explanations" (141). There appears to be a large amount of causal imperialism in XAI - most attempts at XAI make use of causal explanations exclusively, assuming that anything other than a causal explanation is a fictionalization akin to the rationalizations described in section 2.1. Indeed, the bar for explanation under these conditions is so high that some authors have advocated for abandoning the project of developing explainable models entirely, opting instead only for models that are interpretable (Rudin 2019). There are simpler models that exist that are interpretable, such as decision trees, but they are generally less effective than more complex black box

models. The tradeoff with these models is that a causal explanation can be more readily derived when a model is interpretable, because a pattern of counterfactual dependence within the model is easier to discover.

Given their complexity, a causal account of explanation that successfully explains DNNs is likely to be impossible because a pattern of counterfactual dependence cannot be located. The extremely high number of nodes in a DNN, each with an associated weight, is not human parsable. A complete account of causal relationships among nodes will also be non-parsable by humans. AI that is non-interpretable will necessarily also be non-explainable under causal accounts, because to say that a system is non-interpretable is to say that a pattern of counterfactual dependence cannot be established for that system. This follows directly from the definition of non-interpretability. A non-interpretable system is a black box system; when the inner workings of a system are unknown, the causal relationships between that system's components cannot be established. Given the failure of causal accounts in the development of XAI, non-causal accounts of explanation should be explored instead.

The criticisms of attention as explanation from Jain & Wallace and Serrano & Smith implicitly make use of an interventionst account of causal explanation similar to that proposed by Woodward (2003). Because the criticisms of attention as explanation attempt to establish the existence of empirically verifiable causal patterns that hold between the explanandum and those factors without which it would not have occurred, it fits within Woodward's framework. Woodward explains that "an intervention can be thought of as an idealized experimental manipulation which changes C 'surgically' in such a way that any change in E, should it occur, will occur only 'through' the change in C and not via some other route" (Woodward 2018, 119).

In order to determine the existence of causal relationships between variables in a system of variables, the relevant variables are subject to manipulation. Successful explanations, on this account, require that targeted manipulations of relevant system components cause changes in the output of that system when the system output is the explanandum. If manipulations of these parts cause changes to the system's output, the core elements of an explanation are already present. Because the critics of attention as explanation were able to modify seemingly relevant variables without changing the system output, they concluded that deriving an explanation from attention is inappropriate.

The criticisms of attention as explanation implicitly appealed to a view similar to the interventionist account of explanation, but one without a requirement that some variables in the system be held invariant such that the interventions on the system are surgical. Following this requirement ensures that the explanation which is eventually generated can't be superseded by another more plausible explanation related to variables which were not controlled for. In the social sciences, for example, a study of the effects of diet on longevity that does not control for income is likely to be tainted by many spurrious connections between variables that are better explained by the relationship between income and longevity than between diet and longevity. Without holding the extraneous variables invariant, the appropriate pattern of counterfactual dependence cannot be established. The absence of this requirement in the criticisms of attention as explanation may account for the results of these experiments: the discovery of nonsensical alternative explanations derived through the same means, which allowed the researchers to cast doubt on both sets of explanations. The situation does not improve significantly when surgical intervention is used; the problem with applying this approach

to a DNN is that the number of interconnected nodes is so great that engaging in a surgical intervention on any one particular node is likely to be impossible as its value cannot be disentangled from the values of each other node. When making this explicit and taking this requirement into consideration, the outcome is the same - attention is not explanation - but for a different reason. In this case attention is not explanation because under the interventionist framework, it is impossible to engage in surgical intervention on a DNN, and it is thus impossible to find a pattern of counterfactual dependence among the relevant variables within the DNN.

Under the manipulability account of causal explanation, surgical intervention is a method of testing counterfactual conditionals of the form, "if I were to change X in such and such a way, the result would be Y." Actually manipulating the value of X tests the truth of this conditional. Attention is only one part of a larger system of variables. The relevant system in this case is not attention alone, but attention in addition to the DNN itself. While both Jain and Wallace and Serrano and Smith demonstrate the possibility of engaging in surgical intervention on the attention configuration, similar interventions of the remainder of the system are not possible. When surgical intervention is impossible, all counterfactuals are rendered unintelligible since surgical intervention is in one sense merely the testing of a counterfactual conditional. To say that surgical intervention on a given system is impossible is to say that we cannot know the truth of certain counterfactual conditionals about that system.

Of the two case studies explored in section 2.1 and section 2.3, what initially appeared to be the more plausible approach (the use of causal explanations through attention mechanisms in DNNs) now appears as if it may be a dead end. While the use of rationalizations explored in section 2.1 has clear flaws, a factor motivating the

approach, the desire to avoid the messy business of attempting to build causal explanations of DNNs, may have been correct. In the following section I will explore the possibility of applying non-causal explanations to DNNs.

3 Applying Non-Causal Accounts of Explanation to XAI

Both the causal and rationalization approaches to XAI have so far failed to yield good explanations of the decision process happening inside DNNs. The use of rationalizations was an attempt to build psychologically satisfying rather than veridical explanations. The attention example did appear to come closer to an acceptable conclusion. Even if the conclusion was that attention is not explanatory, the discovery of this fact advances the discussion and sets up the possibility for the discovery of other causal explanations in the future. For reasons I discuss below, the use of non-causal explanations is more appropriate for XAI.

The counterfactual theory of explanation (CTE) has causal and non-causal variants. Computer scientists have previously used causal CTE in attempts to build XAI. See, for instance, Wachter et al. (2017) and Sharma et al. (2020) These approaches suffer from many of the same problems identified by computer scientists as discussed in section 2.3.1 and by philosophers as discussed in section 2.4. Alexander Reutlinger (2018) proposes a pluralist extension of the CTE which would allow for both causal and non-causal explanations under the CTE. If it is possible to use a non-causal variant of the CTE to explain DNNs, it might be possible to overcome the objections described in sections 2.3.1

and 2.4.

Mathematical explanation, another candidate category of non-causal explanation of AI, comes, according to Colyvan et al. (2018), in two varieties: intra-mathematical and extra-mathematical. Intra-mathematical explanation is "the explanation of one mathematical fact in terms of other mathematical facts," while extra-mathematical explanation is "the explanation of some physical phenomenon via appeal to mathematical facts" (Colyvan et al. 2018, 232). Extra-mathematical explanation holds great promise for XAI because all DNNs are mathematical. One possible problem is that the relationship between the math used to build AI models and the world is more complicated than, e.g. the relationship between the mathematics used for graph theory when representing the bridges in the city of Königsburg as a graph and the actual city of Königsburg. If an AI classifier is putting images in categories, it can be described and explained in mathematical terms, but the relevant question we seem to want answered isn't about the math, but about the connection between the math and the world. The question of how an AI knows the difference between strawberries and bananas isn't a question limited to its internal mathematical operations because it is also appealing - even if implicitly - to the actual difference between strawberries and bananas. The Seven Bridges of Königsburg problem can be solved with graph theory, but the explanation is still recognizable as representing the actual city of Königsburg. The connection between mathematics and the world in this case is clear, but it is not clear in the case of extra-mathematical explanations of AI.

The potential for the use of models as explanations has been disccused by Bokulich (2011), Batterman & Rice (2014), Morrison (2015), and Potochnik (2017) among others. Model explanations are an exciting possibility for DNNs because DNNs produce models

which are used in decision and classification tasks. If models can serve as explanations, the explanation for deep DNNs could be found in the models they produce (referred to as deep neural models). One major problem with this approach is that with the types of explanatory models discussed in the philosophy of science literature, the model and the phenomena being modeled are different, but in the case of DNNs, the model is the phenomenon that needs to be explained. It is clear from the literature how a model could be explanatory of some external phenomenon, but it is not clear how a model could explain itself. It may be the case that the deep neural model explains the DNN rather than explaining itself, but then the problem of how to explain the model still remains. An explanation of the network that does not also explain the model (which is ultimately responsible for decision and classification tasks) is not enough. It isn't just the DNN which requires an explanation, but the DNN and the model it produces.

4 Conclusion

Because of the high stakes of AI-based decision and classification tasks, explanations of DNNs, deep neural models, and the decisions and classifications they produce are necessary. Computer scientists have attempted to develop explanations of these systems, but their efforts are inadequately grounded in theories of explanation. The study of scientific explanation by the philosophy of science is well suited to this task. non-causal accounts appear to have greater potential to explain DNNs than causal accounts. Non-causal variants of the CTE, extra-mathematical explanations, and model explanations all have potential to provide explanations of DNNs in the future, though more work needs to be done before this is possible. The persistent problems surrounding

explanations of DNNs point to problems with existing accounts of scientific explanation and indicate the necessity for the extension of existing accounts of scientific explanation or the development of new accounts.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." In *Proceedings of ICLR*. 2015.

Batterman, Robert W., and Collin C. Rice. "Minimal Model Explanations." *Philosophy of Science* 81, no. 3 (2014): 349–376. 10.1086/676677.

Bennett, Cynthia L, and Os Keyes. "What is the Point of Fairness? Disability, AI and The Complexity of Justice." In *Workshop on AI Fairness for People with Disabilities at ACM SIGACCESS Conference on Computers and Accessibility*. 2019.

Bokulich, Alisa. "Searching for Non-Causal Explanations in a Sea of Causes." Edited by Alexander Reutlinger and Juha Saatsi. Chap. 7 in *Explanation Beyond Causation*, 141–163. Oxford: Oxford University Press, 2018.

Buckner, Cameron. "Deep Learning: A Philosophical Introduction." *Philosophy Compass* 14, no. 10 (2019). 10.1111/phc3.12625.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. "Predictably unequal? the effects of machine learning on credit markets." *The Effects of Machine Learning on Credit Markets (November 6, 2018)*, 2018.

Harrison, Brent, Upol Ehsan, and Mark O. Riedl. "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations." *CoRR* abs/1702.07826 (2017).

Jain, Sarthak, and Byron C Wallace. "Attention is not Explanation." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 3543–3556. 2019.

Khademi, Aria, and Vasant Honavar. "Algorithmic Bias in Recidivism Prediction: A Causal Perspective." *ArXiv* abs/1911.10640 (2019).

Mark Colyvan, John Cusbert, and Kelvin McQueen. "Two Flavours of Mathematical Explanation." Edited by Alexander Reutlinger and Juha Saatsi. Chap. 11 in *Explanation Beyond Causation*, 231–249. Oxford: Oxford University Press, 2018.

Morrison, Margaret. *Reconstructing Reality: Models, Mathematics, and Simulations*. N.p.: Oup Usa, 2015.

Potochnik, A. *Idealization and the Aims of Science*. N.p.: University of Chicago Press, 2017.

Reutlinger, Alexander. "Extending the Counterfactual Theory of Explanation." Edited by Alexander Reutlinger and Juha Saatsi. Chap. 4 in *Explanation Beyond Causation*, 74–95. Oxford: Oxford University Press, 2018.

Rice, Collin. "Idealized Models, Holistic Distortions, and Universality." *Synthese* 195, no. 6 (2018): 2795–2819.

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1, no. 5 (2019): 206–215.

Serrano, Sofia, and Noah A Smith. "Is Attention Interpretable?" In *Proceedings of ACL*. 2019.

Sharma, Shubham, Jette Henderson, and Joydeep Ghosh. "CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 166–172. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020. 10.1145/3375627.3375812.

Wachter, Sandra, Brent D. Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *CoRR* abs/1711.00399 (2017).

Woodward, James. *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. N.p.: Oxford University Press, 2003.

———. "Some Varieties of Non-Causal Explanation." Edited by Alexander Reutlinger and Juha Saatsi. Chap. 6 in *Explanation Beyond Causation*, 117–137. Oxford: Oxford University Press, 2018.

Yang, Diyi, Aaron Halfaker, Robert E Kraut, and Eduard H Hovy. "Who Did What: Editor Role Identification in Wikipedia." In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*, 446–455. 2016.

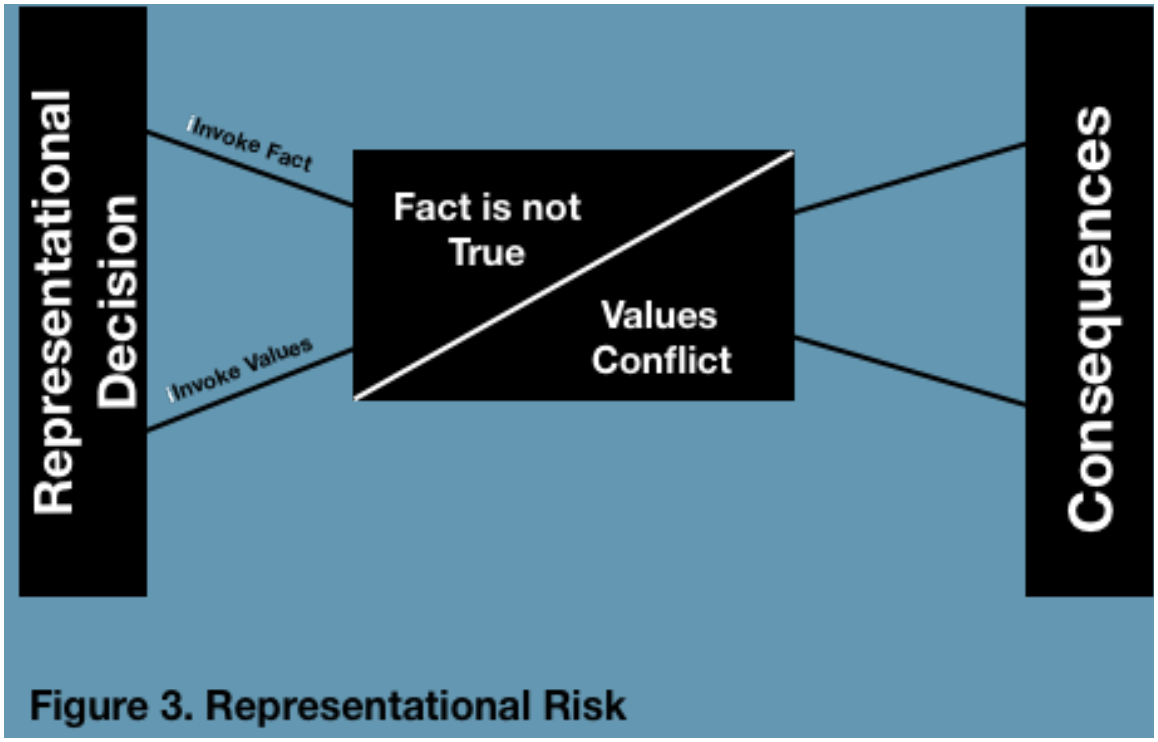


Figure 3. Representational Risk

Understanding & Equivalent Reformulations

Josh Hunt, October 2020

Abstract

Reformulating a scientific theory often leads to a significantly different way of understanding the world. Nevertheless, accounts of both theoretical equivalence and scientific understanding have neglected this important aspect of scientific theorizing. This essay provides a positive account of how reformulating theories changes our understanding. My account simultaneously addresses a serious challenge facing existing accounts of scientific understanding. These accounts have failed to characterize understanding in a way that goes beyond the epistemology of scientific explanation. By focusing on cases where we have differences in understanding without differences in explanation, I show that understanding cannot be reduced to explanation.

1 Introduction

Accounts of theoretical equivalence have neglected an important epistemological question about reformulations: how does reformulating a theory change our understanding of the world? *Prima facie*, improving our understanding is one of the chief intellectual benefits of reformulations. Nevertheless, accounts of theoretical equivalence have focused almost entirely on developing formal and interpretational criteria for when two formulations count as equivalent (Weatherall 2019a). Although no doubt an important question, focusing on it alone misses many other philosophically rich aspects of reformulation.

The burgeoning literature on scientific understanding would seem to be a natural home for characterizing how reformulations improve understanding. However, existing accounts of scientific understanding do not provide a clear answer. These accounts tend to focus on *competing* rather than *compatible* explanations, investigating how the best explanation provides understanding. This strategy neglects how equivalent formulations of the *same explanation* can provide different understandings. To address these gaps, I will show how theoretically equivalent formulations can change our understanding of the world.

Harkening back to Hempel, Kitcher, and Salmon, the *received view* of understanding holds that understanding why a phenomenon occurs simply amounts to grasping a correct explanation of that phenomenon (Strevens 2013; Khalifa 2017, 16ff). Many recent accounts of understanding have decried this picture as overly simplistic, arguing that genuine understanding goes well beyond grasping an explanation (Grimm 2010; Hills 2016; Newman 2017; de Regt 2017). Nevertheless, these critics of the received view still maintain a close connection between explanation and understanding, which Khalifa (2012, 2013, 2015) has exploited to systematically undermine their more expansive accounts. Defending what I'll call *explanationism*, Khalifa (2017) has argued that all philosophical accounts of understanding-why straightforwardly reduce to the epistemology of scientific explanation. Explanationism thereby poses a serious challenge to accounts of scientific understanding that seek to go beyond the traditional received view.

Here, I argue that we can refute explanationism by considering theoretically equivalent formulations. By definition, theoretically equivalent formulations agree completely on the way the world is, thereby describing the exact same state of affairs. Moreover, philosophers

typically adopt an *ontic conception* of explanation, wherein explanations themselves correspond to states of affairs or propositions, e.g. the reasons why an event occurs.[1] By agreeing on the way the world is, equivalent formulations *ipso facto* provide the same explanations. Nonetheless, they can differ radically in the understandings that they provide. Thus, concerning many phenomena, theoretically equivalent formulations do not differ *qua* explanation, even as they differ *qua* understanding. These differences in understanding—without concomitant explanatory differences—make a separate account of understanding necessary.

Section 2 develops Khalifa's challenge for existing accounts of scientific understanding, showing how they reduce to accounts of explanation. I focus in particular on how Khalifa problematizes both skills-based accounts of understanding and a different strategy developed by Lipton (2009) that foreshadows my own. Section 3 demonstrates that theoretically equivalent formulations provide a large class of cases that meet Khalifa's challenge. In these cases, we have differences in understanding-why without differences in explanation. In Section 4, I introduce and defend *conceptualism* as a positive account of these differences in understanding. Conceptualism characterizes how these differences arise from differences in the presentation and organization of explanatory information. Although not a complete account of understanding, conceptualism can be adjoined with existing accounts to both meet Khalifa's challenge and accommodate reformulations. Section 5 considers and rebuts an objection to my use of theoretically equivalent formulations.

2 The challenge from explanationism

Traditional accounts of explanation defend a deflationary stance toward understanding. According to Khalifa, "on the old view, if understanding was not merely psychological afterglow, it was nevertheless redundant, being replaceable by explanatory concepts without loss" (2012, 17). Explanationism encapsulates this deflationary position:

> *Explanationism*: all philosophically significant aspects of understanding-why are encompassed by an appropriately detailed account of the epistemology of scientific explanation.[2]

Importantly, even non-deflationary accounts of scientific understanding must adopt some account of scientific explanation. Then, given whatever account of explanation is adopted, explanationism demands an argument that understanding-why does not reduce to claims about (this kind of) explanation. For this reason, explanationism is dialectically most effective when married with explanatory pluralism (Khalifa 2017, 8).[3] Then, no matter which account(s) of explanation is ultimately correct, explanationism challenges non-deflationary accounts of understanding on their own terms.

Khalifa defends explanationism by developing a detailed account of the epistemology of scientific explanation, which he calls the *explanation-knowledge-science* (EKS) model. According to this framework, an agent improves their understanding why p provided that they

[1] For the ontic conception, see Salmon (1998 [1984], 325), Strevens (2008, 6), Craver (2014), and Skow (2016).

[2] In earlier work, Khalifa refers to this position as the *explanatory model of understanding* (2012, 17). Khalifa (2017, 85) uses "explanationism" in a narrower sense aimed at showing how objectual understanding can be reduced to explanatory understanding, ultimately defending what he calls "quasi-explanationism." For convenience, I simplify this more cumbersome terminology.

[3] Khalifa (2012, 19) claims that explanationism is compatible with explanatory monism, but only if the requisite unified theory of explanation accommodates all typical cases of explanation. It is not clear that such a theory exists.

either (i) gain a more complete grasp of *p*'s *explanatory nexus* or (ii) their grasp of this explanatory nexus more closely resembles *scientific knowledge* (Khalifa 2017, 14). Khalifa defines the *explanatory nexus* as the "totality of explanatory information about *p*," which includes all correct explanations of *p* and the relations between these explanations (2017, 6). I will return to the explanatory nexus in Section 3, arguing that knowledge of this nexus does not exhaust differences in understanding-why. Turning to *scientific knowledge*, Khalifa argues that this requires learning a correct explanation through a process of *scientific explanatory evaluation* (SEEing).[4] Scientific explanatory evaluation involves a three-step process of 1) considering plausible potential explanations, 2) comparing these potential explanations, and 3) deciding how to rank these potential explanations with respect to approximate truth (or at least saving the phenomena) (Khalifa 2017, 12-13). Khalifa uses this ordinary process of SEEing to deflate many anti-explanationist accounts of understanding.

To date, the main anti-explanationist strategy has been to argue that understanding-why involves special skills or abilities. Provided that these skills go beyond what's required for explaining or possessing knowledge-why, explanationism would be refuted.[5] Versions of this *skills-based* strategy include skills for grasping counterfactual information (Grimm 2010, 2014), "cognitive control" over providing and manipulating explanations (Hills 2016), and inferential skills used in making certain kinds of models (Newman 2013, 2017). de Regt has provided one of the most sustained defenses of the skills-based strategy, arguing that understanding involves the ability to make qualitative predictions using an intelligible theory that explains the phenomenon (de Regt and Dieks 2005; de Regt 2009a, 2017).

Khalifa's criticism of Grimm provides the most succinct illustration of explanationism in action. Khalifa argues that Grimm's (2010) account of understanding makes no advance over Woodward's (2003) account of explanation. According to Grimm, understanding is an ability to predict how changing one variable changes another variable, *ceteris paribus* (2010, 340-41). Yet, as Khalifa notes—and Grimm acknowledges (2010, 341; 2014, 339)—this kind of understanding is closely related to Woodward's analysis of "what-if-things-had-been-different questions." Hence, this kind of counterfactual reasoning ability is clearly part of scientific explanatory evaluation (SEEing). We already deploy counterfactual reasoning in considering and comparing alternative explanations, and explaining already involves the ability to answer these what-if questions (Khalifa 2017, 71, 74). Khalifa's response is easily generalized: if all that a theory of understanding adds is referencing a cognitive ability to use an explanation, then a theory of explanation can make the same move without modification.[6]

Another obvious anti-explanationist strategy would involve identifying cases of scientific understanding in the absence of an explanation. Such cases would, at first glance, show that accounts of explanation miss something about understanding. Undertaking precisely this strategy, Lipton (2009) considers a number of cases where we seemingly acquire the cognitive benefits of explanations without actually providing explanations. These cognitive benefits include knowledge of causes, necessity, possibility, and unification (2009, 44). Against the received view, Lipton identifies understanding not as "having an explanation," but rather with "the cognitive benefits that an explanation provides" (2009, 43). Notice that this still

[4]Khalifa also requires that this belief-forming process be *safe*, i.e. sufficiently unlikely to lead to false beliefs.

[5]Some epistemologists have pursued other strategies, arguing that objectual understanding either does not reduce to understanding-why or else that some forms of objectual understanding do not even require explanatory understanding. Khalifa responds at length to these approaches (2017, 80ff).

[6]Khalifa (2012) applies this strategy to criticize de Regt and Dieks (2005) and de Regt (2009a, 2009b) in detail. Against Hills, Khalifa argues that her necessary conditions for understanding are either irrelevant for enhancing understanding or else they are captured by the EKS model (2017, 70-72). He responds to Newman in his (2015).

maintains a close connection between understanding and explanation.

Khalifa (2013) exploits this connection to argue that Lipton's strategy makes no fundamental advance over the explanation literature. Systematically examining each of Lipton's examples, Khalifa shows that whenever there is understanding through a non-explanation, there is an explanation that provides *that* understanding *and more*. This leads to what Khalifa calls "*explanatory idealism*" about understanding, which holds that "other modes of understanding ought to be assessed by how well they replicate the understanding provided by knowledge of a good and correct explanation" (2013, 162). Thus, a suitably detailed account of scientific explanation would provide the same insights about understanding that Lipton defends. In this way, explanation functions as the "*ideal* of understanding" (Khalifa 2013, 162). More recently, Khalifa (2017) has recast part of his criticism as what he calls the "right track objection." According to this objection, Lipton's examples involve agents who merely have a kind of "proto-understanding," wherein they are on the right track to acquiring an explanation and thereby understanding-why.

In the remainder of this essay, I defend a strategy that avoids Khalifa's objections against existing accounts of scientific understanding. My strategy succeeds where others fail for two reasons. First, I do not rely on positing any special abilities unique to understanding, so Khalifa's challenge from SEEing does not apply. Secondly, the examples I consider provide understanding through the same explanatory information, so explanatory idealism does not apply either.

3 Intellectual differences without explanatory differences

To refute explanationism, it suffices to identify differences in understanding-why between two presentations of the *same* explanation, since these appeal—*ipso facto*—to the same explanatory information. In such cases, understanding-why still arises from an explanation, but non-explanatory differences account for the corresponding differences in understanding. The features we ascribe to "understanding-why" and to "explanation" then truly come apart. For convenience, I will refer to differences in understanding as *intellectual differences*. This section aims to show that, *pace* explanationism, we can have intellectual differences without concomitant explanatory differences.

To forestall any hopes of a piecemeal explanationist rebuttal, my argument requires a sufficiently large class of examples stemming from scientific practice. As we will see, the recent literature on theoretical equivalence provides a rich set of cases, spanning many parts of physics. Nevertheless, some might worry that these mathematical reformulations are too isolated or special to be indicative of scientific understanding in general. Hence, it is worthwhile to also consider a more common aspect of scientific practice: diagrammatic reformulations. I will consider both cases in turn, illustrating each with a paradigmatic example.[7] Importantly, my argument does not apply to cases of different but *complementary* explanations, such as Salmon's example of causal-mechanical vs. unificationist explanations of a balloon moving forward upon takeoff in an airplane (Salmon 1998, 73; de Regt 2017, 77). Such complementary explanations appeal to different explanatory information and are hence genuinely different explanations. Khalifa's EKS model of understanding accommodates such cases since they reference different parts of the explanatory nexus (2017, 25).

By definition, *theoretically equivalent formulations* express the same scientific theory,

[7]Reformulations of symmetry arguments provide another class of examples. See Hunt (forthcoming) for details.

agreeing exactly on the way the world is (or could be). Philosophers have defended a few different characterizations of theoretical equivalence, including definitional equivalence (Glymour 1971), model isomorphism (North 2009), and categorical equivalence (Halvorson 2016; Weatherall 2016; Barrett 2019). These accounts all seek to formalize the intuition that two formulations are theoretically equivalent if and only if they are mutually inter-translatable and empirically equivalent. Mutual inter-translatability requires that any thing expressed in one formulation can be expressed in the other without loss of physically significant information. Empirical equivalence requires that the formulations agree on all physically possible measurable consequences.

Recent defenses of categorical equivalence have shown it to be the most fruitful criterion for theoretical equivalence. It successfully formalizes a number of philosophically and scientifically plausible cases of theoretically equivalent formulations.[8] Five prominent examples include Lagrangian and Hamiltonian formulations of classical mechanics (Barrett 2019), standard and geometrized formulations of Newtonian gravity theories (Weatherall 2016), Lorentzian manifold and Einstein algebra formulations of general relativity (Rosenstock et al. 2015), Faraday tensor and 4-vector potential formulations of classical electromagnetism (Weatherall 2016), and principal bundle and holonomy formulations of Yang–Mills gauge theories (Rosenstock and Weatherall 2016). Here, then, is a varied class of cases that collectively pose a substantive problem for explanationism.

In each of these cases, I contend, we have intellectual differences without corresponding explanatory differences. Each formulation provides a different understanding than its equivalent counterpart for at least the following simple reason: understanding one does not entail understanding the other (and indeed, showing that they are equivalent requires nontrivial insights). For instance, understanding a phenomenon via Lagrangian mechanics does not entail an understanding of that same phenomenon using Hamiltonian mechanics. Thus, Lagrangian understanding-why differs from Hamiltonian understanding-why, even though both involve grasping the same explanation. The lack of explanatory differences follows from categorical equivalence, which entails that we can inter-translate models of one formulation into models of the other without losing any information.[9] In other words, equivalent formulations possess "the same capacities to represent physical situations" (Rosenstock et al. 2015, 315). On the common ontic conception of explanation assumed here, explanatory information itself is a subset of this physical information, so equivalent formulations *a fortiori* represent the same explanatory information. Thus, whenever one formulation provides an explanation, any equivalent formulation provides the same explanation, preserving everything of explanatory significance—but not necessarily of intellectual significance.

Lagrangian and Hamiltonian mechanics provide a simple but detailed illustration of the foregoing points.[10] These equivalent formulations display two main sources of intellectual differences. First, they differ in how they encode the system's dynamics. The Lagrangian formalism uses a *Lagrangian function* $L(q_i, \dot{q}_i, t)$, encoding the dynamics as a function of time t, generalized coordinates q_i, and generalized velocities $\dot{q}_i$.[11] In the Hamiltonian formalism, we perform a variable change from generalized velocities to generalized momenta p_i, yielding the Hamiltonian $H(q_i, p_i, t)$. Despite encoding the same physical information, the Lagrangian

[8]For an introduction see Halvorson (2016, 601) and for technical details Weatherall (2016, 2019b).

[9]For defenses of this claim, see Weatherall (2016, 1083, 1087) and Rosenstock et al. (2015, 314).

[10]Technically—within a subclass of models known as the hyper-regular domain—Barrett (2019) shows that the Lagrangian tangent bundle and Hamiltonian cotangent bundle formulations are equivalent. For ease of exposition, I present their more elementary coordinate-based formalisms. For details see Goldstein et al. (2002).

[11]Here, the index i runs over {1, 2, ..., n}. The "$\dot{v}$" notation indicates a first derivative with respect to time.

and Hamiltonian organize this information differently, as illustrated below. Secondly, the two formulations represent the dynamical laws of evolution (the equations of motion) in dramatically different ways. Whereas the Lagrangian formulation represents these as a set of n-many *2nd-order* differential equations (the Euler–Lagrange equations), the Hamiltonian formulation represents these same equations of motion as a set of $2n$-many *1st-order* differential equations (Hamilton's equations).[12] By reorganizing the equations of motion in this way, the Hamiltonian formulation treats the generalized coordinates q_i and the generalized momenta p_i more symmetrically. This leads to further intellectual differences in some cases, such as the symmetry argument considered next.

A typical explanandum in mechanics concerns the evolution of a classical system such as a pendulum or spinning top. In systems with symmetry, one generalized coordinate—e.g. q_n—is typically *ignorable*, meaning that it does not occur in the Lagrangian or Hamiltonian.[13] The equations of motion then entail that the corresponding conjugate momentum, p_n, is a conserved quantity, i.e. a constant α. It is here that a dramatic intellectual difference occurs between the formulations. Despite p_n being constant, the corresponding generalized velocity $\dot{q}_n$ need not be. Hence, $\dot{q}_n$ still appears in the Lagrangian as a nontrivial variable. A Lagrangian understanding of the system's evolution thereby still requires considering n-many degrees of freedom, despite having an ignorable coordinate. In contrast, the Hamiltonian formalism enables a genuine reduction in the number of degrees of freedom that need to be considered, resulting in a different understanding. Thanks to changing variables from generalized velocities to generalized momenta, the Hamiltonian depends on the latter but not the former. Hence, we can replace p_n in the Hamiltonian with a constant α, and—with the ignorable coordinate q_n also absent—this eliminates an entire degree of freedom from consideration.[14] As Butterfield remarks, this example "illustrates one of mechanics' grand themes: exploiting a symmetry so as to reduce the number of variables needed to treat a problem" (2006, 43). Although not an explanatory difference, this variable reduction demonstrates a difference in how the same explanatory content is organized. This organizational difference results in a different understanding of the system's evolution. Indeed, these kinds of organizational differences ultimately lead to differences in understanding Noether's first theorem—a foundational result connecting continuous symmetries and conserved quantities (Butterfield 2006).

Thanks to their rigorous mutual inter-translatability, categorically equivalent formulations provide the most precise illustration of my argument. However, at a less rigorous level, theoretically equivalent formulations arise whenever we reformulate a theory while keeping its physical content the same. This motivates including at least some instances of diagrammatic reasoning within the class of theoretically equivalent formulations. Although neglected by the literature on theoretical equivalence, diagrammatic reformulations satisfy the same intuitive criteria: mutual inter-translatability and empirical equivalence. They thereby provide another large class of examples where we can have differences in understanding-why without concomitant explanatory differences. Examples of diagrammatic reformulations include Feynman diagrams in particle and condensed matter physics, graphical approaches to the quantum theory of angular momentum (Brink and Satchler 1968), Penrose–Carter diagrams in space-time theories, graph-theoretic approaches to chemistry (Balaban 1985; Trinajstić

[12] In both cases, we require $2n$ initial values to solve these equations.

[13] It is easy to show that a generalized coordinate does not appear in the Lagrangian if and only if it does not appear in the Hamiltonian.

[14] Technically, we replace one of Hamilton's equations with a trivial integral for calculating $\dot{q}_n$.

1992), and diagrams for mechanistic reasoning in biology (Abrahamsen and Bechtel 2015).

To illustrate how diagrammatic reasoning can provide intellectual differences, consider Feynman diagrams in particle physics. Here, the explanandum is typically a scattering amplitude for a particular interaction, explained by calculating terms in a perturbation expansion. Without using Feynman diagrams, we can calculate each term up to a desired order in perturbation theory. This provides one way of understanding the scattering amplitude. Alternatively, we can reorganize this same explanatory information using Feynman diagrams, allowing us to express *connectivity properties* of terms in the perturbation expansion. To calculate the scattering amplitude, it suffices to know the connected terms; the disconnected terms do not contribute.[15] Focusing on connectivity thereby makes it unnecessary to consider a vast number of terms in the perturbation expansion—terms that a brute force calculation would show vanish. In this way, Feynman diagrams lead to a different understanding of scattering amplitudes but without introducing any additional explanatory information.[16]

4 A conceptualist account of understanding

I have argued that a variety of mathematical and diagrammatic reformulations provide intellectual differences without associated explanatory differences. Yet, if not from explanatory differences, whence do these intellectual differences arise? To answer this question, I will introduce and defend *conceptualism*, which claims that intellectual differences result from differences in how explanatory information is organized and presented. These organizational differences lead to differences in *what we need to know* to present explanations, leading to differences in understanding-why. I will consider an objection that conceptualism merely describes how reformulations modify explanatory concepts, with no effect on understanding-why. To rebut this objection, I will argue that nontrivial changes in explanatory concepts necessarily lead to differences in understanding-why.

Conceptualism posits a sufficient condition for differences in understanding-why: reformulating an explanation generates an intellectual difference whenever it changes what we *need to know* or *what suffices to know* to present that explanation. For instance, in shifting from Lagrangian mechanics to Hamiltonian mechanics, we learn that we don't need to know how to represent the system and its dynamics using the Lagrangian and the Euler–Lagrange equations. Knowledge of the Hamiltonian and Hamilton's equations suffices. *Mutatis mutandis*, the same can be said for shifting from Hamiltonian mechanics to Lagrangian mechanics, leading again to a difference in understanding. Similarly, reformulating scattering amplitude explanations using Feynman diagrams teaches us that we don't need to know the disconnected terms in the perturbation expansion: knowledge of the connected terms suffices. For convenience, I will refer to these differences in what-we-need-to-know or what-suffices-to-know as *epistemic dependence relations* (EDRs). Conceptualism claims that when equivalent formulations provide different epistemic dependence relations, they manifest intellectual differences.

To rebuff explanationism, these intellectual differences must be genuine differences in *understanding why* empirical phenomena occur. If instead these intellectual differences con-

[15] A term is connected if there is a path of propagators connecting every pair of source factors and/or vertex factors in the term. For technical background and formal results, see for instance Srednicki (2007, §§8–10) and Lancaster and Blundell (2014, §§16–20, 22, and 24).

[16] de Regt (2017, 251ff) also considers Feynman diagrams to defend his account of understanding. Whereas he focuses on visualization, I focus only on formal features that are independent of human psychology.

cern some other kind of understanding, explanationism is left unscathed. Accordingly, an explanationist might argue that differences in EDRs do not genuinely affect understanding-why. Rather, these differences might merely affect our understanding of the *concepts* used to represent explanations, concepts such as Lagrangians, Hamiltonians, connected diagrams, Lorentzian manifolds, etc.[17] If so, conceptualism would have failed to identify a genuine source of intellectual differences.

Conceptualism agrees with part of this objection: in the first instance, reformulating an explanation changes our understanding *of that explanation*. However, nontrivial changes in understanding an explanation entail differences in understanding-why. Conceptualism reframes this claim as a simple bridge principle:[18]

> *Intellectual bridge principle* (IBP): A nontrivial difference in understanding an explanation of p leads to a different understanding why p.

According to this bridge principle, organizing the same explanatory information differently can lead to a different understanding-why, as we have seen in the case of Lagrangian and Hamiltonian mechanics. Different ways of understanding an explanation are *nontrivial* provided that they are not merely conventional differences in presenting an explanation. Hence, the intellectual bridge principle excludes a large class of *trivial notational variants* from counting as intellectually significant.[19] For instance, uniformly replacing "5" everywhere with "V" in an Arabic numeral system would result in different presentations of many explanations, but these differences would be trivial, rather than intellectually significant. Similarly, recasting an explanation using a left-handed coordinate system rather than a right-handed one would not result in any differences in understanding-why. Although it is difficult to precisely delimit trivial from nontrivial notational variants, my defense of conceptualism requires only the existence of clear cases of nontrivial reformulations, such as those developed in Section 3. In general, conceptualism posits that a difference in epistemic dependence relations is both necessary and sufficient for an intellectually significant difference.[20] Trivial notational variants do not provide different EDRs and hence do not generate intellectual differences.

In response, an explanationist might attempt to reject this bridge principle. However, the IBP follows straightforwardly from the received view of understanding, which explanationism seeks to uphold. Recall that according to the received view, understanding why a phenomenon occurs amounts to grasping an explanation of that phenomenon. Grasping explanations requires that we can represent them, and any way of representing explanations involves concepts. Hence, understanding the relevant explanatory concepts is necessary for understanding-why. Understanding-why is thereby derivative on the way that we have understood this explanation, such as the epistemic dependence relations we have used to present it. Thus, at least some changes in explanatory concepts must lead to concomitant changes in understanding-why. In other words, any account of understanding requires a bridge principle to connect our explanatory concepts with achieving understanding.

With these distinctions in hand, conceptualism straightforwardly identifies the origins of intellectual differences between the equivalent formulations mentioned in Section 3. To take

[17] I adapt this objection from Khalifa (2017, 138), who develops it as a further argument against Lipton (2009).

[18] de Regt similarly argues that understanding a phenomenon necessarily requires being able to understand a theory (2017, 44). However, I disagree with de Regt that understanding a theory is always pragmatic and contextual.

[19] Grammatically, "intellectually significant" is analogous to "explanatorily significant." It characterizes differences that matter for understanding.

[20] Reasons of space prevent a detailed defense of this claim, which I defend elsewhere.

one example, the Einstein algebra formalism is markedly different from the standard formulation of general relativity. It teaches us that we don't need to know the standard Lorentzian manifold and metric concepts to provide explanations in general relativity. Instead, we can reorganize all of the relevant explanatory information using algebraic notions, as Geroch (1972) has argued. Since this reformulation changes what we need to know to present explanations, it is not a trivial notational variant of the standard formulation. It thereby satisfies the intellectual bridge principle, leading to a different understanding-why for phenomena explained by general relativity.

By itself, conceptualism does not provide a full-fledged account of scientific understanding. Instead, it illuminates an important facet of understanding that has been neglected in the literature. Due to its minimal commitments, conceptualism can be adjoined with existing accounts of understanding, particularly those allied against explanationism. Although compatible with skills-based accounts of understanding, conceptualism does not assume any special role for skills or abilities. The key insight behind my position is that how a theory-formulation organizes explanatory information matters for understanding. Scientific agents perform no more special a role than grasping this organizational structure. For these reasons, my position is not susceptible to the explanationist strategy against skills-based accounts considered in Section 2. Likewise, since conceptualism focuses on how recasting explanations changes understanding, it does not succumb to Khalifa's objections to Lipton's (2009) *understanding without explanation* proposal.

5 An objection against theoretical equivalence

Prima facie, one strategy remains available to an explanationist: they can reject my argument in Section 3 that theoretically equivalent formulations provide the *same* explanation. Instead, they might argue that in such cases, one formulation takes explanatory priority. There are at least two candidate sources of explanatory priority. First, one formulation might be physically privileged. For instance, Curiel (2014) privileges Lagrangian mechanics for allegedly encoding the kinematic constraints of classical systems. Secondly, one formulation might be more fundamental or joint-carving than another. This metaphysical difference would presumably entail a corresponding explanatory difference, wherein the more fundamental formulation provides a better explanation (Sider 2011, 61). Differences in joint-carving or perfectly natural properties would then be part of the explanatory nexus. For instance, North (2009) argues that Hamiltonian mechanics is more fundamental than Lagrangian mechanics.

However, this objection sits uneasily within the broader dialectical strategy of explanationism. Recall from Section 2 that to problematize multifarious accounts of understanding, explanationism adopts a form of explanatory pluralism. Otherwise, it is all too easy to designate some aspects of explanation (e.g. the causal-mechanical ones) as genuinely explanatory while other aspects (such as unification) are seen as mattering for understanding but not explanation. Furthermore, adopting explanatory pluralism seems to require a modicum of ontological pluralism as well (Khalifa 2017, 7). This is because different models of explanation take different ontological features as necessary for providing explanations, as shown in recent debates over causal vs. noncausal explanations (Lange 2017).

Hence, insofar as explanationism requires both explanatory and ontological pluralism, it cannot preclude the interpretation of theoretically equivalent formulations adopted in Section 3. It must allow philosophers to interpret cases of theoretically equivalent formulations

as being just that: genuinely equivalent both physically and metaphysically.[21] If explanationists instead adopt a single account of explanation, they will be unable to systematically recast all purported differences in understanding as explanatory differences. The explanationist is thus caught on the horns of a dilemma. Either they renounce explanatory pluralism and thereby fail to systematically deflate skills-based accounts of understanding, or they maintain pluralism and thereby allow that theoretically equivalent formulations provide the same explanation but different understandings.

6 Conclusion

I have argued that theoretically equivalent formulations provide a clear counterexample to explanationism. Whereas explanationism holds that all intellectual differences arise from explanatory differences, equivalent formulations show that some differences in understanding-why do not reduce to explanatory differences. To accommodate these intellectual differences, I have proposed *conceptualism*. Conceptualism argues that understanding-why involves not only the explanatory content that we have understood, but also the way that we have understood it. In particular, it claims that equivalent formulations manifest intellectual differences whenever they provide different *epistemic dependence relations*. These are differences in what we need to know or what suffices to know to provide an explanation. By characterizing how reformulations change understanding, conceptualism addresses complementary lacunae in current accounts of both scientific understanding and theoretical equivalence. In this way, conceptualism supplements existing anti-explanationist accounts of scientific understanding. By adopting conceptualism, these accounts can forestall the challenge from explanationism and genuinely go beyond the epistemology of scientific explanation.

References

Abrahamsen, Adele, and William Bechtel. 2015. "Diagrams as tools for scientific reasoning". *Review of Philosophy and Psychology* 6.

Balaban, Alexandru. 1985. "Applications of graph theory in chemistry". *Journal of Chemical Information and Computer Sciences* 25 (3).

Barrett, Thomas William. 2019. "Equivalent and Inequivalent Formulations of Classical Mechanics". *British Journal for Philosophy of Science* 70 (4).

Brink, D. M., and G. R. Satchler. 1968. "Graphical methods in angular momentum". In *Angular momentum*, 2nd. Oxford: Clarendon.

Butterfield, J. N. 2006. "On symmetry and conserved quantities in classical mechanics". In *Physical theory and its interpretation: essays in honor of Jeffrey Bub*, ed. by William Demopoulos and Itamar Pitowsky. Dordrecht: Springer.

Craver, Carl F. 2014. "The ontic account of scientific explanation". In *Explanation in the special sciences: The case of biology and history*, ed. by M.I. Kaiser et al. Springer.

[21] As Rosenstock et al. note, "it seems far more philosophically interesting to recognize that the world may admit of such different, but equally good, descriptions than to argue about which approach is primary" (2015, 315–16).

Curiel, Erik. 2014. "Classical Mechanics Is Lagrangian; It Is Not Hamiltonian". *The British Journal for the Philosophy of Science* 65.

de Regt, Henk. 2009a. "The epistemic value of understanding". *Philosophy of Science* 76.

— . 2009b. "Understanding and scientific explanation". In *Scientific understanding: philosophical perspectives*. University of Pittsburgh.

— . 2017. *Understanding Scientific Understanding*. Oxford.

de Regt, Henk, and Dennis Dieks. 2005. "A contextual approach to scientific understanding". *Synthese* 144.

Geroch, Robert. 1972. "Einstein algebras". *Communications in Mathematical Physics* 26.

Glymour, Clark. 1971. "Theoretical realism and theoretical equivalence". In *Proceedings of the Biennial Meeting of the Philosophy of Science Association*.

Goldstein, Herbert, et al. 2002. *Classical mechanics*. 3rd. San Francisco: Addison Wesley.

Grimm, Stephen R. 2010. "The goal of explanation". *Studies in History and Philosophy of Science* 41.

— . 2014. "Understanding as knowledge of causes". In *Virtue epistemology naturalized*. Springer.

Halvorson, Hans. 2016. "Scientific Theories". In *The Oxford Handbook of Philosophy of Science*, ed. by Paul Humphreys. Oxford.

Hills, Alison. 2016. "Understanding why". *Noûs* 50 (4).

Hunt, Josh. Forthcoming. "Epistemic dependence and understanding: reformulating through symmetry". *The British Journal for the Philosophy of Science*.

Khalifa, Kareem. 2015. "EMU defended: reply to Newman (2014)". *European journal for philosophy of science* 5 (3).

— . 2012. "Inaugurating understanding or repackaging explanation?" *Philosophy of Science* 79 (1).

— . 2013. "The role of explanation in understanding". *The British Journal for the Philosophy of Science* 64.

— . 2017. *Understanding, explanation, and scientific knowledge*. Cambridge.

Lancaster, Tom, and Stephen J Blundell. 2014. *Quantum field theory for the gifted amateur*. Oxford.

Lange, Marc. 2017. *Because without cause: Non-causal explanations in science and mathematics*. Oxford.

Lipton, Peter. 2009. "Understanding without explanation". In *Scientific understanding: philosophical perspectives*. University of Pittsburgh.

Newman, Mark P. 2013. "Refining the inferential model of scientific understanding". *International studies in the philosophy of science* 27 (2).

— . 2017. "Theoretical Understanding in Science". *The British Journal for the Philosophy of Science* 68 (2).

North, Jill. 2009. "The 'Structure' of Physics: A Case Study". *The Journal of Philosophy* 106.

Rosenstock, Sarita, and James Owen Weatherall. 2016. "A categorical equivalence between generalized holonomy maps on a connected manifold and principal connections on bundles over that manifold". *Journal of Mathematical Physics* 57 (10).

Rosenstock, Sarita, et al. 2015. "On Einstein algebras and relativistic spacetimes". *Studies in History and Philosophy of Modern Physics* 52.

Salmon, Wesley C. 1998. "Scientific explanation: Causation and unification". In *Causality and explanation*. Oxford.

— . 1998 [1984]. "Scientific explanation: Three basic conceptions". In *Causality and explanation*. Oxford.

Sider, Theodore. 2011. *Writing the Book of the World*. Oxford.

Skow, Bradford. 2016. *Reasons why*. Oxford.

Srednicki, Mark. 2007. *Quantum field theory*. Cambridge.

Strevens, Michael. 2008. *Depth: An account of scientific explanation*. Harvard.

— . 2013. "No understanding without explanation". *Studies in history and philosophy of science Part A* 44 (3).

Trinajstić, Nenad. 1992. *Chemical graph theory*. 2nd. Boca Raton: CRC Press.

Weatherall, James Owen. 2016. "Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent?" *Erkenntnis* 81 (5).

— . 2019a. "Theoretical equivalence in physics: Part 1". *Philosophy Compass* 14.

— . 2019b. "Theoretical equivalence in physics: Part 2". *Philosophy Compass* 14.

Woodward, James. 2003. *Making things happen: a theory of causal explanation*. New York: Oxford.

How Much Change is Too Much Change? Rethinking the Reasons Behind the Lack of Reception to Brouwer's Intuitionism

[version accepted for presentation at PSA 2020; please don't cite or quote without author's permission]

Kati Kish Bar-On

The Cohn Institute for the History and Philosophy of Science and Ideas, Tel Aviv University

katikish@gmail.com

Abstract

The paper analyzes Brouwer's intuitionistic attempt to reform mathematics through the prism of Leo Corry's philosophical model of "body" and "image" of knowledge. Such an analysis sheds new light on the question of whether Brouwer's intuitionism could at all be attractive to broader groups of mathematicians. It focuses on three characteristics that are unique to Bouwer's reformation attempt and suggests that when considered together, they combine to provide a more complex understanding of the reasons behind the lack of reception to Brouwer's intuitionism than any of the three can offer alone.

1. Introduction

Brouwer's intuitionistic program was an intriguing attempt to reform the foundations of mathematics and was probably the most controversial one within the contours of the foundational debate during the 1920s (van Stigt 1990; Hesseling 2003). Historians and philosophers of mathematics have tried to account for the reasons why Brouwer's intuitionism did not prevail. Some associate the demise of Brouwer's intuitionism with his dismissal from the editorial board of the *Mathematische Annalen* in 1928 (van Atten 2004). Others suggest that the lack of reception derived from technical difficulties within Brouwer's mathematical arguments (Epple 2000) or due to his awkward and too-technical style of writing (van Dalen 2013).

In the following pages, I wish to discuss a specific aspect of the question of whether Brouwer's far-reaching intuitionistic program could at all be attractive to broader groups of mathematicians. In order to do that, I would like to consider the story of Brouwer's intuitionism in light of Leo Corry's model of image and body of knowledge, and alongside Corry's compelling analysis of Van der Waerden's *Moderne Algebra*, which created new knowledge from mathematical notions that already existed. I intend to focus on three significant differences between the stories of Brouwer and Van der Waerden: on their different motivations for change, on the scope of the change, and on the implications of using familiar mathematical concepts (as opposed to introducing completely new notions). The variations between the two stories, I shall argue, offer a new perspective on the lack of reception to

Brouwer's intuitionism, that is owed not only to technical difficulties within the theory but to a combination of a deep philosophical motivation (with which mathematicians were less sympathetic), a too comprehensive reformation, and contradicting new mathematical concepts.

The terminology of 'image of knowledge' and 'body of knowledge' is borrowed from Yehuda Elkana's work. According to Elkana, the 'body of knowledge' is where the research is being done; thus, it consists of different theories, concepts, and mechanisms (Elkana 1978, 315). The 'images of knowledge' governs particular aspects of scientific activity that the 'body of knowledge' does not address, like: sources of knowledge, the legitimization of knowledge, the audience of knowledge, and relatedness to prevailing norms and ideologies. Building on Elkana's theory, the process of scientific progress can be described as engaging with two different types of questions: the first addresses the methods used in the process of making a discovery or forming a new theory, and the second addresses the guiding principles and normative boundaries of the discipline itself.

Unlike other disciplines, mathematics is uniquely endowed with a special interconnection between its body and image of knowledge. The reflexive aspect of mathematics enables it to examine the nature of the discipline itself by applying the same framework that is used in everyday methodological practice[1] (Corry 1989). Some mathematical theories can be easily

[1] Consider, for example, proof theory. No other discipline has a dedicated practical doctrine about how its methods should be properly done.

classified into one of the two realms, while other arguments may encompass an aspect of both. Upon considering past attempts to transform a constituting mathematical framework, a series of questions arise regarding the place of such revolutionizing theories: do they evolve from the body of knowledge, the image of knowledge, or both? Is there a specific path or order for changes to occur? Does a shift in one layer must proceed the other?

The historian Leo Corry suggests that there is not only one direction in which mathematical transitions can occur (Corry 2001). As a case study, Corry examines the structural image of a specific mathematical discipline, namely, algebra, by analyzing van der Waerden's *Moderne Algebra*, which presents the body of algebraic knowledge as deriving from a single unified perspective, and all the relevant results in the field are achieved using similar concepts and methods (Corry 2001, 172). The systematic study of different varieties of algebra through a common approach is what Corry calls a structural image of algebra, and whereas the transition to a new structural image in the case of van der Waerden's *Moderne Algebra* was enabled due to changes in the body of knowledge, it does not imply that this is mandatory. Thus, transitions between images of knowledge are unique and distinct processes from transitions in the body of knowledge. Corry perceives the body and image of knowledge as organically interconnected domains in the history of a discipline, but he does not regard their relation as a cause and effect.

In the case of van der Waerden's *Moderne Algebra*, the newly proposed image had firm roots in the then-current body of knowledge. Though the textbook presents an original perspective

regarding the algebraic structure, it uses as cornerstones several mathematical notions such as groups, fields, and ideals that have already been introduced to the mathematical community, and it builds upon already developed theories of renown algebraists (such as Emmy Noether and Ernst Steinitz). Van der Waerden took mathematical concepts (such as Isomorphism) that were previously defined separately for different mathematical notions (such as groups, rings, or fields) and showed that they could be a-priori defined for each algebraic system (Schlote 2005; Corry 2001). The mathematical entities van der Waerden discussed were familiar and acceptable within the mathematical discourse; the novelty he introduced lied in the relations between them.

The notions van der Waerden applied in Moderne Algebra did not appear there for the first time: the concept of 'group' was already found in algebra textbooks from 1866, and the notions of ideals and fields were introduced by Dedekind in 1871. Brouwer, on the other hand, introduced new, original concepts and theories that were meant to replace the old, classical, non-constructive ones.

To suggest an alternative to the set-theoretical notions of a class of numbers, Brouwer employed two intuitionistic analogs: ‘species’ and ‘spread.’ A species is a property that mathematical objects can have, and objects with this property are called the elements of the species. A spread is a collection of sequences called the nodes of the spread and is defined by a ‘spread function’ which performs a decidable procedure on finite sequences (Troelstra 1969; Dummett 1977). Another new concept that Brouwer introduced was ‘choice sequences,’ also

called ‘infinitely proceeding sequences’ (Troelstra 1977; van der Hoeven & Moerdijk 1984). Such sequences need be neither law-like (that is, governed by computable recipes for generating terms) nor even fully determinate in advance. Nothing about the future course of the sequence may be known, other than the fact that its terms are freely and independently chosen[2].

The problematic aspect of spread, species, and choice sequences (that are only taken here as representatives among several other new intuitionistic concepts Brouwer had introduced[3]), does not lie solely in their novelty. It is entwined with Brouwer’s motive to develop these new intuitionistic concepts, namely, his philosophical views that put philosophy before mathematics (and not the other way around). Brouwer was willing to forego significant parts of mathematics in order to refrain from the paradoxes of set theory, but for mathematicians, the scope of the change was far too comprehensive. Practicing mathematicians wish to solve problems at the core of the discipline, not to contemplate philosophical conundrums. Here lies another significant difference between Brouwer’s and van der Waerden’s stories: due to Brouwer’s philosophical views, the whole foundational basis of mathematics had to change.

[2] Brouwer permitted restrictions imposed by a spread law, but nothing beyond that.

[3] From these notions, together with the new definition of the natural numbers as mental constructs, Brouwer goes on to formulate additional intuitionistic concepts and theories such as bar theorem, fan, and fan theorem (Dummett 1977, van der Hoeven & Moerdijk 1984).

Brouwer's intention was primarily to reform the foundations of mathematics, but van der Waerden's agenda was utterly different. Even tough *Moderne Algebra* turned out to be an influential book that had a significant impact on algebra as a discipline, 'reformation' was not what van der Waerden had in mind. To take seriously the question of whether intuitionism could at all appeal to a broader mathematical audience, we must consider the combination of several differences between the two stories. It is not only the use of familiar or non-familiar mathematical notions, but also the philosophical motives for change (or lack thereof), and the scope of the change that shape the way mathematicians read and respond to new ideas. In order to gain a better understanding of the differences between Brouwer's and van der Waerden's motives and scope of change, let us explore the contours of Brouwer's intuitionism.

2. The scope of change as suggested by Brouwer's intuitionistic program

Brouwer's intuitionism holds that the existence of an object is equivalent to the possibility of its construction in one's mind. There is an important philosophical distinction between objects like finite numbers and constructively given denumerable sets, which are objects that we finite beings can intuitively grasp, and the Cantorian collection of all real numbers, which is an infinite entity that exceeds our limited grasp. Brouwer regarded the former entities as 'finished' or 'finish-able' while the latter are 'unfinished.' A 'finished' set is produced by a recognizable process (that is, a process that one can construct), yielding some legitimate grasp

of the object with all its parts (that is, that the parts are 'determined' by the initial grasp). An 'unfinished' collection is one that we cannot grasp in a way that suffices to determine all its parts (Brouwer 1952; Posy 2008).

Throughout his dissertation, Brouwer uses this differentiation to confront Cantor's perception of infinity. Brouwer accepts ω-sequences as legitimate mathematical objects since it is a sequence of discrete elements that are generated by a countably ordered process (Brouwer 1912, 85-86), but it is the only infinite object he accepted (Brouwer 1907, 142–143).

Brouwer addressed the set of real numbers as 'denumerably unfinished' from a negative perspective, pointing out that given a denumerable subset, we can straightaway find an element of the continuum that is not in the given subset, but there is no positive existence claim to support it. Hence, he proclaimed Cantor's second number class and any ranked order of increasing cardinalities as illegitimate mathematical objects, a mere "expression for a known intention" (Brouwer 1907, 148).

As for the intuition of the continuum itself, Brouwer firmly believed that we have an intuitive grasp of the continuum as a whole (Brouwer 1907, 8-9, 62). Thus, a continuum that is constructed out of a set of independently given points (like the Cantorian continuum) cannot be considered a legitimate mathematical entity. No set of points can exhaust the continuum since, in Brouwer's view, it is a unity in its own right (Posy 2005). Building on from this concept of the continuum and his notion of infinity, Brouwer postulated a separate form of

intuition that delivers the continuum as a whole and generates the 'mathematics of the continuum,' thereby creating a new body of mathematical knowledge.

By virtue of Brouwer's new concept of a potential infinity, core notions like the principle of excluded middle and the concept of negation are deemed unacceptable in Brouwer's intuitionism. The principle of excluded middle can only be used as a reliable tool in finite systems where each object of the set can be examined (in principle) by means of a finite process. Within a finite system, one can eventually determine whether there is a member of the set with the property A or that every member of the set lacks the property A. However, in infinite systems it is no longer possible to examine every object of the set (not even in principle); thus, even if one never finds a member of the set with the property A, it does not prove that every member of the set lacks the property A (Brouwer 1908, 1918).

Together with the restricting concept of infinity and his demand that mathematical objects must be constructed, Brouwer introduced a new image of mathematical knowledge, which he considered as the only proper way to do mathematics. As a result of such changes, the idea of mathematical truth and its relation to the provability and refutability of a mathematical statement was redefined: in the newly proposed intuitionistic theory knowing that a statement P is true means having proof of it. Otherwise stated, to assert that a statement P is true is to claim that P can be proved; to negate P is to claim that P is refutable (i.e., that a counterexample exists), but it does not imply that "not P" is provable (Brouwer 1912; Heyting 1966; Sundholm and van Atten 2008). One of the many implications from such an utter

reformation was that proofs of mathematical existence by contradiction ceased to be a legitimate technique within the discipline, inducing a change both to the image and to the body of knowledge.

Moreover, Brouwer's newly proposed image excluded several central mathematical theories, and extensively altered other widespread mathematical concepts. Among some of the classical theories Brouwer was willing to eschew, was Zermelo's axiom of choice, that was referred to by Hilbert as constituting "a general logical principle which, even for the first elements of mathematical inference, is indispensable" (Moore 1982, 253), emphasizing the considerable differences between the new and the existing bodies of knowledge.

3. Differences in the process of creating new knowledge

According to Corry, the innovative aspect of van der Waerden's book was that it created new and significant mathematical knowledge without introducing any new mathematical entities, theories, or concepts. Van der Waerden took mathematical concepts and elements that were developed within specific, different mathematical contexts and realized that within the framework of algebra, a variation of the same elements could be axiomatically defined, studied, and brought together into a new conceptual organization of the discipline. Such mathematical concepts were "different varieties of a same species ("varieties" and "species"

understood here in a "biological," not mathematical term), namely, different kinds of algebraic structures" (Corry 2001, 176).

The type of knowledge created in van der Waerden's *Moderne Algebra* can be regarded, to some extent, as a continuation of the already existing mathematical body of knowledge. Brouwer's story was rather different; the extensive scope of reformation Brouwer imposed on the prevailing body of mathematical knowledge, including the restricting concept of infinity and the intuitionistic notion of the continuum as a whole, made intuitionism altogether incomparable with classical mathematics. The intuitionistic approach is not merely a restriction of classical reasoning; it contradicts classical mathematics in a fundamental way (Iemhoff 2019).

More than it was a new mathematical approach to the foundational problem of mathematics, Brouwer's Intuitionism was, first and foremost, a philosophy of mathematics. Tracing back to his 1907 Ph.D. dissertation, Brouwer intended to work out his ideas in the philosophy of mathematics, rather than to describe various views on the foundations of mathematics (van Dalen 1981). In a letter from Brouwer to his supervisor, Diederick Korteweg, Brouwer wrote that he is glad he is finally able to use mathematics in order to support his criticism of the value and usefulness of language and logic[4].

[4] As documented in a letter from Brouwer to Korteweg from September 1906 (taken from Van Dalen 1981, 5).

Brouwer's dissertation consisted of three chapters: 'The construction of mathematics,' 'Mathematics and experience,' and 'Mathematics and logic.' In the first chapter, Brouwer constructs mathematics from the natural numbers to the negative, the rational, and the irrational numbers and introduces the continuum as an ever unfinished (Brouwer 1907, 44-52; Van Dalen 1999). The goal of Brouwer's second chapter is to improve Kant's view of the a priori, as in Brouwer's opinion the only a priori element in science is the intuition of time since the creation of the image of space is a free act of the intellect and as such cannot be part of the a priori. In the third and last chapter Brouwer touches the two themes that will become the most central issues in the foundational debate: the principle of excluded middle and mathematical existence, and directs his criticism towards Hilbert's idea of securing the foundations of mathematics by consistency proof (Brouwer 1907, 176).

Korteweg's main criticism of Brouwer's dissertation was directed towards the second chapter, as he firmly objected to the idea of philosophical mathematics as a scientific topic for a dissertation. Korteweg read parts of Brouwer's book Life, Art, and Mysticism, but he expected Brouwer to separate between his philosophical and mathematical work[5]. However, in

[5] It should be noted that despite his criticism of Brouwer's philosophical views, Korteweg was one of Brouwer's most prominent advocates. He was a firm believer in Brouwer's mathematical abilities and did everything in his power to secure him an academic position. Throughout his attempt to get Brouwer elected to the Academy of Sciences in 1910, Korteweg

Brouwer's case, philosophy was "the basic ingredient that made the mathematics work" (van Dalen 2013, 86), and he had no intention in restricting his philosophical activity to leisure hours. Korteweg was concerned with the reception of Brouwer's work in the faculty, specifically with the philosophical and moral views Brouwer presented as part of his second chapter. He expressed his misgivings in a letter to Brouwer, where he stated:

> After receiving your letter I have again considered whether I could accept it as it is now. But really Brouwer, this won't do. A kind of pessimistic and mystic philosophy of life has been woven into it that is no longer mathematics, and has also nothing to do with the foundations of mathematics. It may here and there have coalesced in your mind with mathematics, but that is wholly subjective. One can in that respect totally differ with you, and yet completely share your views on the foundations of mathematics. I am convinced that every supervisor, young or old, sharing or not sharing your philosophy of life, would object to its incorporation in a mathematical dissertation. In my opinion your dissertation can only gain by removing it. It now gives it a character of bizarreness which can only harm it. (van Dalen 2013, 92-93)

---

even approached leading international mathematicians such as Hilbert and Poincaré in order to get their recommendations for a membership to a "gifted and exceptional scholar" such as Brouwer (van Dalen 2013).

Korteweg's remarks on Brouwer's dissertation stress out the extent to which Brouwer's work deviated from the acceptable norm of a standard mathematical dissertation. Korteweg suspected that Brouwer's philosophical incentive to reform mathematics might not appeal to a wide mathematical audience, mainly as it contradicted familiar notions and stood against the theories developed by momentous mathematicians like David Hilbert. Some changes are too far-reaching for mathematicians to endure, and even the glamorous promise of consistency and non-paradoxical foundations is not enough to make them give up the methods and theories they use doing everyday mathematics.

Eventually, Brouwer revised the second chapter, leaving some of the problematic philosophical parts out of it. However, Brouwer maintained his philosophical views and continued to develop intuitionism primarily as a philosophy of mathematics that entailed a massive reformation to the foundations of the discipline rather than an extension or a continuation of classical mathematics. The role of intuition in Brouwer's mathematics is a means of introducing new mathematical structures, not 'different kind of the same species' structures as in van der Waerden's work. As David Hesseling puts it, Brouwer "started from his own ideas and looked for mathematics that fitted in, instead of working the other way around" (Hesseling 2003, 35).

4. Concluding Remarks

The current paper suggests that three factors played a significant role in the lack of reception to Brouwer's intuitionism: the first is Brouwer's philosophical agenda to reform mathematics, which was foreign and unrelatable in the eyes of most mathematicians, including Brouwer's Ph.D. supervisor, Korteweg. The second is the extensive scope of change Brouwer had imposed on all aspects of mathematics, thereby excluding major theories and acceptable proofs from the legitimate body of mathematical knowledge. The third element is the introduction of new concepts and theories that were meant to replace the non-constructive ones, a move that only further characterized intuitionism as an isolated theory, deprived of any foothold in current mathematical practices. Each factor is entwined with and explains the other two factors: Brouwer's philosophical motivation is the reason behind the massive scope of reformation he suggested, and every new concept he introduced is rooted in his philosophical views of what mathematics actually is and how it should be practiced. Taken alone, each factor is necessary but not sufficient in the attempt to explain why the intuitionistic program was not able to attract broader groups of mathematicians. However, all three factors together offer a more comprehensive picture of the lack of reception to Brouwer's intuitionism.

Among the mathematicians who did embrace Brouwer's intuitionism for a short period was Hermann Weyl, who was deeply influenced by philosophers such as Fichte and Husserl (Scholz 2000, Feferman 1998) and tried to develop his philosophical view of mathematics in his monograph *Das Kontinuum* (Weyl 1918). However, Weyl was quite a unique scholar

among mathematicians and physicists that not only contributed to his own fields of research but also engaged with philosophical questions about the foundations of mathematics and the nature of mathematical entities (Weyl 1949). Most mathematicians restricted their areas of expertise to the discipline of mathematics and did not find a necessary connection between practicing mathematics and doing philosophy. Even Brouwer's prominent student, Arend Heyting, was ambivalent in regards to the inseparable connection between mathematics and philosophy that Brouwer had imposed.

Heyting's intuitionism and Brouwer's intuitionism were quite different: while Brouwer insisted on detaching intuitionism from any axiomatic method, Heyting took Brouwer's intuitionistic ideas and expressed them using a formalistic approach. Heyting's formalization comprised intuitionistic propositional and predicate logic, arithmetic, and analysis, all together in one big system (Heyting 1930; 1980). While his formalization of analysis did not derive from its classical counterpart (thus it was somewhat overlooked within the foundational debate), the parts concerned with logic and arithmetic were subsystems of their classical counterparts (expect from the principle of excluded middle, which was excluded from Heyting's theory), and were extensively discussed (van Atten 2017).

Heyting's intuitionism reached a wide mathematical audience and continued to develop over the following decades (see: Gentzen 1935; Heyting 1966; Kleene 1952; Myhill 1966; Vesley 1980). Was the reaction to Heyting's intuitionism a result of the differences between Heyting's intuitionism and Brouwer's intuitionism? Was it only the formalization of

intuitionism that made Heyting's ideas more approachable to working mathematicians, or was it also the sense of detachment from Brouwer's philosophical stances that is evident in Heyting's approach to intuitionistic mathematics? There is a conflicting view regarding the way Heyting addressed Brouwer's philosophical considerations. Albeit Heyting quoted Brouwer's remarks on the relation between mathematics and logic, Heyting also claimed that philosophy is not necessary in order to understand intuitionistic mathematics, and some of Brouwer's most significant concepts (such as consciousness and mind) play no role in Heyting's approach (Heyting 1974; Placek 1999). Unlike Brouwer, Heyting did not attempt to justify the intuitionistic revision philosophically, nor to suggest that philosophical assumptions are inherent in intuitionistic mathematics. As opposed to Brouwer's viewpoint, Heyting argued that intuitionism is much simpler than any philosophy and that it would be better for the sake of intuitionism to eliminate any philosophical (metaphysical as well as epistemological) premises. As Heyting put it:

> The only philosophical thesis of mathematical intuitionism is that no philosophy is needed to understand mathematics. (Heyting 1974, 79).

The superiority of intuitionistic mathematics over classical mathematics, according to Heyting, derives from the former being free from any metaphysical or philosophical assumption. Therefore, it appears that even Brouwer's most devoted student deviated from Brouwer's philosophical positions and regarded them as alien and even irrelevant to Brouwer's intuitionistic program. What caused Heyting to abandon Brouwer's philosophical approach

but to continue his intuitionistic pursue? Is it possible that Brouwer's philosophical views combined with the massive reformation he suggested and the introduction of new concepts demanded too much mathematical compromises, even from a faithful disciple like Heyting? Can the examination of Heyting's reception of Brouwer's intuitionistic mathematics while discarding Brouwer's philosophy shed yet another light on the question of why other mathematicians, less devoted to Brouwer, were unwilling to accept Brouwer's intuitionistic program? The current paper sets the stage for exploring these questions and provides a prolific ground to start from, in its attempt to present a more inclusive perspective on how certain developments in mathematics prevailed whereas others did not.

References:

Brouwer, Luitzen Egbertus Jan. 1907. "On the Foundations of Mathematics", Ph.D. Diss., University of Amsterdam.

——. 1908. "The Unreliability of Logical Principles." In *L.E.J Brouwer Collected Works I – Philosophy and Foundations of Mathematics*, ed. Arend Heyting, 105-11. Amsterdam: North-Holland.

——. 1912. "Intuitionism and Formalism." In *L.E.J Brouwer Collected Works I – Philosophy and Foundations of Mathematics*, ed. Arend Heyting, 123-38. Amsterdam: North-Holland.

——. 1918. "Founding Set Theory Independently of the Principle of the Excluded Middle." *KNAW Verhandelingen*, 5:1–43.

——. 1952. "Historical background, principles and methods of Intuitionism." In *L.E.J Brouwer Collected Works I – Philosophy and Foundations of Mathematics*, ed. Arend Heyting, 508-15. Amsterdam: North-Holland.

Corry, Leo. 1989. "Linearity and Reflexivity in the Growth of Mathematical Knowledge." *Science in Context* 3(2): 409-40.

——. 2001. "Mathematical Structures from Hilbert to Bourbaki: The Evolution of an Image of Mathematics." In: *Changing Images in Mathematics: From the French Revolution to the New*

Millennium, eds. Umberto Bottazzini, Amy Dahan-Dalmédico, 167-86. Harwood Academic Publishers.

Dummett, Michael. 1977. *Elements of Intuitionism*. Clarendon Press, Oxford, England.

Elkana, Yehuda. 1978. "Two-Tier Thinking: Philosophical Realism and Historical Relativism." *Social Studies of Science,* 8:309-26. SAGE, London and Beverly Hills.

Epple, Moritz. 2000. "Did Brouwer's intuitionistic analysis satisfy its own epistemological standards?" In: *Proof theory: History and philosophical significance*, eds. V. F. Hendricks et al., 153-78. Dordrecht: Kluwer.

Feferman, Solomon. 1998. *In the Light of Logic*. New York: Oxford University Press.

Gentzen, Gerhard. 1935. "Untersuchungen Über das logische Schliessen." *Mathematsche Zeitschrift* 39:176–210.

Hesseling, David E. 2003. *Gnomes in the Fog: The Reception of Brouwer's intuitionism in the 1920s*. Basel, Boston, Berlin: Birkhu-ser Verlag.

Heyting, Arend. [1930] 1998. "Die formalen Regeln der intuitionistischen Logik." In *From Brouwer to Hilbert: The Debate on the Foundations of Mathematics in the 1920s,* ed. Paolo Mancosu, 311-27. Oxford: Oxford University Press.

——. 1966. *Intuitionism: An introduction*. Amsterdam: North-Holland.

——. 1974. "Intuitionistic views on the nature of mathematics." *Synthese*, 27(1–2): 79–91.

——. 1980. *Collected Papers*. Amsterdam: North-Holland.

Iemhoff, Rosalie. 2019. "Intuitionism in the Philosophy of Mathematics." *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition).

Kleene, Stephen Cole. 1952. *Introduction to Metamathematics*. Princeton: Van Nostrand.

Moore, Gregory H. 1982. “Zermelo's axiom of choice: Its origins, development, and influence.” *Studies in the history of mathematics and physical sciences*, vol. 8. Springer-Verlag, New York, Heidelberg, and Berlin.

Myhill, John. 1966. “Notes Towards An Axiomatization of Intuitionistic Analysis.” *Logique et Analyse*, 9(35–36): 280–97.

Placek Tomasz. 1999. “Heyting’s Arguments.” *Mathematical Intuitionism and Intersubjectivity*, 279: 103-46. Springer, Dordrecht.

Posy, Carl. 2005. “Intuitionism and philosophy”. In: *The Oxford Handbook of Philosophy of Mathematics and Logic*, ed. Shapiro, Stewart, 318-55. Oxford University Press.

——. 2008. “Brouwerian infinity.” In: *One Hundred Years of Intuitionism (1907–2007)*, eds. van Atten M., Boldini P., Bourdeau M., Heinzmann G., 21-36, Publications of the Henri Poincaré Archives, Birkhäuser Basel.

Schlote, K.-H. 2005. “B.L. van der Waerden, moderne algebra, first edition (1930–1931).” In: *Landmark Writings in Western Mathematics 1640-1940*, ed. Grattan-Guinness, Ivor, 901-16. Elsevier.

Scholz, Erhard. 2000. "Herman Weyl on the Concept of Continuum." In: *Proof theory: History and philosophical significance,* ed. V. F. Hendricks et al., 195-220. Dordrecht: Kluwer.

Sundholm, Goran, & van Atten, Mark. 2008. “The proper explanation of intuitionistic logic: on Brouwer’s demonstration of the Bar Theorem.” In: *One Hundred Years of Intuitionism* (1907–2007), eds. van Atten M., Boldini P., Bourdeau M., Heinzmann G., 60-77. Birkhäuser Basel.

Troelstra, Anne Sjerp. 1969. *Principles of intuitionism*. Springer, Berlin.

——. 1977. *Choice sequences*. Clarendon Press.

van Atten, Mark. 2004. *On Brouwer*. Wadsworth/Thomson Learning, Belmont.

——. 2017. "The Development of Intuitionistic Logic." *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), ed. Edward N. Zalta.

van Dalen, Dirk, ed. 1981. *Brouwer's Cambridge Lectures on Intuitionism,* Cambridge University Press, Cambridge.

——. 1999. *Mystic, Geometer, and Intuitionist: The Life of L.E.J. Brouwer*. Volume I: The dawning Revolution, Clarendon Press, Oxford.

——. 2013. *L.E.J. Brouwer – Topologist, Intuitionist, Philosopher: How Mathematics Is Rooted in Life.* London: Springer-Verlag.

van der Hoeven, Gerrit, & Moerdijk, Ieke. 1984. "On choice sequences determined by spreads." *Journal of Symbolic Logic* 49:908–16.

van Stigt, Walter. 1990. *Studies in the History and Philosophy of Mathematics: Brouwer's Intuitionism*. North-Holland, Amsterdam.

Vesley, Richard. 1980. "Intuitionistic Analysis: the Search for Axiomatization and Understanding." In *The Kleene Symposium (Studies in Logic and the Foundations of Mathematics)*, eds. J. Barwise, H. J. Keisler, and K. Kunen, 317–31. Amsterdam: North-Holland.

Weyl, Hermann. 1918. *The Continuum: A Critical Examination of the Foundation of Analysis.* Thomas Jefferson University Press.

——. 1949. *Philosophy of Mathematics and Natural Science*. Princeton: Princeton University Press.

REFLEXIVITY, FUNCTIONAL REFERENCE, AND MODULARITY: ALTERNATIVE TARGETS FOR LANGUAGE ORIGINS

TRAVIS LACROIX

Abstract. Researchers in language origins typically try to explain how compositional communication might evolve to bridge the gap between animal communication and natural language. However, as an explanatory target, compositionality has been shown to be problematic for a gradualist approach to the evolution of language. In this paper, I suggest that *reflexivity* provides an apt and plausible alternative target which does not succumb to the problems that compositionality faces. I further explain how *proto*-reflexivity, which depends upon functional reference, gives rise to complex communication systems via modular composition.

Keywords — reflexivity, language origins, explanatory targets, functional reference, modular composition, compositionality, animal communication

1. Introduction

Communication is ubiquitous in nature: every taxon that has been investigated displays some form of communication system (Kight et al., 2013). However, *linguistic* communication—i.e., natural language—is (or at least is often taken to be) unique to humans. This raises the question; *how did language evolve?* That is, how did rich linguistic communication systems like the ones we see in humans evolve *out of* simpler non-linguistic systems of communication? This is an inherently difficult question due to a lack of direct evidence—language does not fossilise, and we cannot observe the actual precursors of human language in, e.g., extinct hominin ancestors.

Nonetheless, work on language origins has blossomed in recent decades. New data, increasingly sophisticated techniques and technologies, and productive interdisciplinary research have helped foster the development of subtle models of language evolution. This is achieved using a multi-component approach to understand the mechanisms underlying language and how they might have evolved (Fitch,

Mila - Québec Artificial Intelligence Institute
Département d'informatique et de recherche opérationnelle, Université de Montréal
E-mail address: lacroixt@mila.quebec.
Date: Draft of June 15, 2020.

2017). For example, *comparative methods* in evolutionary biology start by breaking down a complex trait into multiple subcomponent mechanisms or features (Fitch, 2017; Martinez, 2018). We can then examine the presence or absence of traits, in phylogenetic terms, to infer facts about whether some particular trait common to several species is a *homologue* or an *analogue*. *Computer simulations* further provide a concrete and explicit way to test hypotheses (Cangelosi and Parisi, 2002), furnishing a *how-possibly* explanation of the sort that is common in evolutionary biology (Resnik, 1991). However, the plausibility of these results requires figuring empirical evidence from relevant fields—in the case of language origins, this includes evidence from biology, linguistics, animal communication, neuroscience, and more.

The most common feature of natural language that is appealed to as a gap-bridging explanatory target is compositionality (and related features like hierarchy and recursion). The idea is that if we could explain how compositional communication can evolve out of non-compositional communication, we would have taken great strides in explaining how language evolved. However, this is problematic insofar as (1) compositionality, in an evolutionary context, proffers asymmetric benefits for senders and receivers of signals, and researchers have not maintained adequate sensitivity to this role-asymmetry (LaCroix, 2020a); (2) there is no empirical evidence for proto-compositional communication as a precursor to natural language insofar as the oft-cited evidence is more likely homologous to human-level linguistic compositionality than analogous (LaCroix, 2019a); and (3) there is no gradualist explanation of compositionality, insofar as this is a binary property of language (Berwick and Chomsky, 2011; LaCroix, 2020b).

In this paper, I propose that *reflexivity*—the ability to use language to talk about language—provides an apt and plausible alternative explanatory target for language-origins research. I further explain how *proto*-reflexivity, which depends upon *functional reference*, gives rise to complex communication systems via modular composition. I argue that reflexivity does not succumb to the problems that compositionality faces since (1) role asymmetries are accounted for by the underlying mechanism of functional reference, (2) there exists empirical evidence of plausible precursors to reflexivity in nature, (3) the precursors of reflexivity are graded. Finally, reflexivity allows for rich compositional structures that have been shown to give rise to genuinely compositional syntax.

2. Proto-reflexivity, functional reference, and their evolutionary precursors

Communication is a unique evolutionary system in the following sense. Once a group of individuals has learned some simple communication convention, those learned behaviours may be used to influence future communicative behaviour,

thereby affecting future communication conventions. This may give rise to a feedback loop, wherein more complex communication, in turn, is used to influence future communicative behaviours which are even more sophisticated.

When faced with a novel context, individuals can always learn a brand-new disposition from scratch. However, in some cases, it may be more advantageous or more efficient to utilise a pre-evolved disposition. When individuals take advantage of pre-evolved *communicative* dispositions to thereby influence future communication, this is a form of *proto-reflexivity*. Such an ability is an evolutionary precursor to the reflexivity of natural languages, wherein one can use language to talk *about* language.

Proto-reflexivity depends primarily upon *functional reference*, which has been the subject of much empirical and theoretical work in animal communication (Sievers and Gruber, 2016). Functional reference is so-called because it is meant to evoke the idea of reference in language without being equivalent to reference in the way that words refer. So, the ability to refer functionally is an evolutionary precursor to the ability to refer linguistically. Signals are functionally referential if they are 'elicited by a special class of stimuli and capable of causing behaviours adaptive to such stimuli in the absence of contextual cues' (Scarantino, 2013, 1006).[1] They are therefore *context-specific* for the signaller to produce, and *stimulus-independent* for the receiver to understand. This can be defined formally, as in Definition 2.1.

> **Definition 2.1:** (*Strong*) *Functional Reference*
> A token of type X *functionally refers* to a token of type Y just in case the following two criteria are jointly satisfied:
>
> (1) *Production Criterion*: Xs are reliably caused (only/mostly) by Ys;
> (2) *Perception Criterion*: Xs presentations reliably cause responses adaptive to Ys in the absence of Ys and any other contextual cues.

For example, vervet monkey (*Chlorocebus pygerythrus*) alarm calls are suggested (Seyfarth et al., 1980) to be functionally referential since the presence of an eagle (Y) reliably causes an eagle alarm call (X), satisfying the production criterion. Furthermore, the presentation of an eagle alarm call (X) reliably causes recipients to hide in the bush (an adaptive response to the presence of an eagle, Y), satisfying the perception criterion. Playback experiments suggest that these responses occur in the absence of other contextual cues.

Female Diana monkeys (*Cercopithecus diana*) elicit alarm calls upon viewing a predator first-hand and respond to alarm calls of male Diana monkeys by repeating the call. Zuberbühler et al. (1999) perform playback experiments of various pairs

[1] See also Macedonia and Evans (1993).

of stimuli—a matching pair consists of an alarm call followed by the sound of the predator to which the call functionally refers; a mismatched pair consists of an alarm call followed by the sound of a predator to which the call does not functionally refer. In each case, pairs of stimuli are separated by five minutes of silence. In the experiment, the female monkeys displayed less concern upon hearing, e.g., the characteristic shriek of an eagle five minutes after the eagle alarm call—the former conveys no new information. However, they showed significant concern upon hearing a characteristic leopard growl five minutes after hearing the eagle alarm call. The conclusion is that alarm calls do not just serve to trigger (behaviourally or deterministically) an evasive response: individuals have an 'idea'—what Hurford (2007) terms a 'proto-concept'—of the relevant predator in mind for at least five minutes following the initial alarm call.

We might worry about the strength of Definition 2.1 since, for example, aggression signals may functionally refer to future aggressive behaviour, though it perhaps seems strange to say they are caused by it. We can weaken this by indexing to a context and replacing causation with correlation, as in Definition 2.2 (Scarantino, 2013):

> **Definition 2.2:** (*Weak*) *Functional Reference*
> A token of type X **in context C** *functionally refers* to a token of type Y just in case the following two criteria are jointly satisfied:
>
> (1) *Contextual Information Criterion*: Xs **in context** C are **correlated** with Ys (weakly or strongly);
> (2) *Contextual Perception Criterion*: Xs presentations **in context** C reliably cause responses adaptive to Ys in the absence of Ys.

This is information-theoretic because X carries information about Y just in case Xs and Ys are correlated.[2] The intuition is that the signal and the functional referent must correlate enough to make responding to the signal in ways that are adaptive to the referent evolutionarily advantageous.

Functional reference, and therefore proto-reflexivity, minimally requires several communicative precursors, including arbitrariness, specialisation, semanticity, discreteness, and displacement (Hockett, 1960). *Arbitrariness* requires that there is no 'natural' connection between a linguistic form and its meaning; this contrasts with *iconic* signals where there is a similarity between the form of a sign and its meaning—e.g., onomatopoeia in natural language. *Specialisation* requires that the signal produced is intended for communication, and not because of another

[2]This dovetails nicely with the role that information transfer plays in studies of animal communication (see Stegmann (2013), though cf. Dawkins and Krebs (1978)), as well as theoretical work in philosophy on meaning as informational content; see Skyrms (2010a,b).

behaviour; this contrasts with *cues*, which are a byproduct of some other (non-communicative) process—e.g., the presence of CO_2 transfers information about the location of a mammal, though exhalation of CO_2 did not evolve for this purpose. *Semanticity* requires that there is a relationship between a signal and its meaning. However, these three features of communication are early-evolving abilities that are common to mammals generally. *Discreteness* means that signals are perceived categorically, as opposed to continuously; this feature is present in primates generally. Finally, *displacement* is the ability to talk about things that are not present in the immediate environment.

Consider a situation where individuals coordinate upon a communication convention, like in a simple signalling game (Lewis, 1969; Skyrms, 2010a). In this case, the messages may functionally refer to the states of the world—as in the vervet monkey alarm call system. Now, suppose that this signalling situation occurs in a pre-evolved context. Suppose further that there is a novel context in which individuals must learn a new communication system. In some cases, the output of the novel signalling context may be an appropriate *input* for the pre-evolved signalling context (Barrett and Skyrms, 2017; LaCroix, 2020c); see Figure 1.

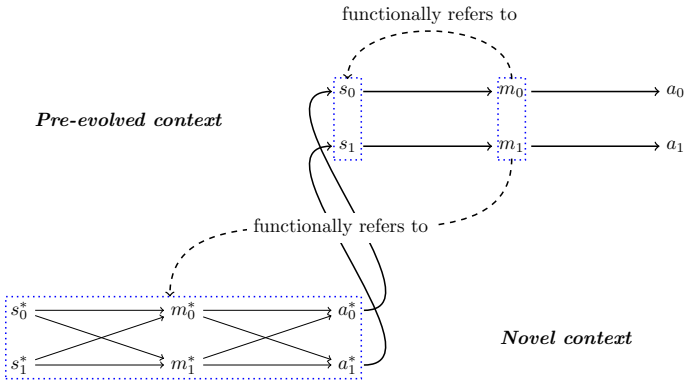


FIGURE 1

However, signals functionally refer to states in the pre-evolved context, and the states are just the output of the communication system in the novel context; so, messages come to functionally refer to the communication system itself, in a way that is *proto*-reflexive: they functionally refer to a *communication* context as a whole, rather than linguistic symbols themselves. In such a situation, *discrete*, *arbitrary*, and *meaningful* signals, which are *specialised* for communicative purposes, come to (functionally) refer to something abstract, in a sense, and so *displaced* from the immediate environment.

How might such a property or ability evolve? This happens by way of *modular composition* and related processes. Various processes of this sort may include appropriation or template transfer, analogical reasoning, or genuine modular composition.

3. Modular composition and related processes

3.1. **Transfer (of) Learning.** The simplest way of evolving new strategies from old strategies is *appropriation*. This process, minimally, requires the following. First, the agents must have evolved a disposition for a particular context. The agents then face a novel context, where the prior disposition *just happens* to be appropriate—though this may not be known at the outset. This novel context may be relevantly similar, but non-trivially distinct, from the original context. Appropriation then consists in applying the prior strategy to the novel context. It may be that the agent happens, by chance, to try something pre-evolved when faced with a novel context. The appropriateness of the pre-evolved strategy may determine a sufficiently beneficial reward such that, when faced with this same context again, the agent learns quickly (even by simple reinforcement) to perform the old action. This simple form of appropriation is sometimes called *transfer (of) learning*.[3]

This allows for flexibility of behaviour in problem-solving, via the ability to generalise learned rules to novel contexts. There is good evidence that many species of new- and old-world monkeys, as well as great apes, are capable of transfer; however, prosimians are not (Rumbaugh, 1970, 1971, 1995; Rumbaugh and Pate, 1984a,b; Bonte et al., 2014). One example of transfer learning in nonhuman animals is an extension of classification tasks, involving 'reversal learning'. Here, an animal is trained to associate a particular stimulus with a reward. Once the agent exhibits some degree of success, the relation between the stimulus and the reward reverses, so the agent must replace the prior association with the *opposite* association. If the animal can quickly reverse its associations, it is assumed that successful performance is based on a *concept* of oppositeness. On the other hand, if the new association takes as long or longer to be learned, no such application of conceptual understanding may be attributed to the agent.[4]

Minimally, transfer learning requires *only* that an agent try prior strategies. Successful strategies may be learned via simple reinforcement, or they may be discovered via a more sophisticated trial-and-error. When salience is present—e.g., the physical properties of a new predator being saliently similar to an old predator—the

[3]See, e.g., Ellis (1965); Schunk (2004); Pugh and Bergin (2006); Hung (2013).

[4]Hurford (2007) argues that reversal learning experiments do not merely highlight an ability to apply the relation of oppositeness between a source and a target context; instead, the agent 'seems to be keeping its old mental representation (concept) of the general class of stimuli acquired in the first training regime and relating the new set to that acquired concept' (25).

new strategy may be implemented immediately; however, this is a more sophisticated version of transfer learning, which requires a concept of *analogical similarity*.

3.2. **Analogical Reasoning.** The most common way of testing analogical reasoning ability is with a set of analogy problems known as *relational matching-to-sample* (RMTS) tasks.[5] This experimental task involves showing the agent a sample set, which consists of two or more objects that are either identical or non-identical. The agent is then shown two comparison sets, which contain novel objects—one of which involves identity, and the other of which involves non-identity. To be successful, the agent must choose the comparison set which matches the sample set.

In this case, the analogy between various stimuli requires a concept of SAME versus DIFFERENT. As with transfer learning, there is some evidence that nonhuman animals can utilise analogical reasoning. Despite prior belief to the contrary (Thompson and Oden, 2000), it has been shown experimentally that some apes (importantly, chimpanzees) can perform these tasks easily. Other apes and very few old-world monkeys can perform these tasks, but only after extensive training. In each case, symbolic training results in better performance, implying a relationship between cognition and linguistic ability.[6]

Noting and taking advantage of analogy is more cognitively complex than simple transfer. Increasing complexity again, we arrive at a full concept of *modular composition*.

3.3. **Modular Composition.** Finally, modular composition itself varies in complexity, but the most complex forms are supposed to be unique to humans and to depend upon language. Spelke (2003) suggests that humans and other animals are endowed with early-developing, core systems of knowledge, called 'modules'. However, these core systems are limited in several ways. First, they are *domain-specific*, since these modules represent only a subset of entities in the surroundings of the agent. Second, they are *task-specific*, since they inform only a subset of the repertoire of the agent's actions and cognitive processes. Third, they are (at least relatively) *encapsulated*, since there is a restriction on the flow of information into and out of a module. Finally, modules are (at least relatively) isolated from one another, since they do not readily combine (Spelke, 2003, 291).[7]

[5]See Skinner (1950); Blough (1959); Ferster (1960).
[6]See, e.g., Skinner (1950); Blough (1959); Ferster (1960); Fagot et al. (2001); Wasserman et al. (2001); Katz et al. (2002); Flemming et al. (2011).
[7]See also Fodor (1983, 1984, 2000); Sherry and Schacter (1987); Sperber (1994); Coltheart (1999); Sperber (2002); Carruthers (2002); Barrett and Kurzban (2006); Shettleworth (2012); Robbins (2017).

Many core cognitive capacities that are available to (and were once thought to be unique to) humans are also available to nonhuman animals (Spelke, 2003).[8] Therefore, humans, but also nonhuman animals, have early-developing core knowledge systems, which allow for a broad range of intelligent behaviour and cognitive capacities; and, in many cases, these same core systems enable nonhuman animals to outperform human infants in similar tasks. Thus, core systems alone do not account for uniquely human cognitive capacities. Spelke (2003) suggests that human cognitive capacities depend on core knowledge systems, which are shared by other animals, *and* on a uniquely human combinatorial ability for conjoining these representations to create new systems of knowledge. Furthermore, she suggests that the latter capacity is made possible by natural language, which provides the medium for combining the representations delivered by core knowledge systems (305). Specifically, it is the *compositional* nature of natural language, which gives rise to uniquely flexible human cognition, on her account.

The basic communicative abilities that give rise to human linguistic capacities are shared with many other species; however, the ability to produce and interpret recursive structures is uniquely human (Hauser et al., 2002). If we assume that the human capacity for language can be decomposed into a set of well-defined mechanisms that interact via interfaces, then we can begin to examine how such interfaces between individual components may 'hook up' in the first place. In essence, this is the concept of modular composition as it is described in Barrett and Skyrms (2017). Modular composition ties together explanations of complexity in communicative, cognitive, and social structures.

4. Reflexivity as an explanatory target

Researchers typically propose evolutionary theories that explain how compositionality arose, moving from a one-word stage (simple signalling), to a two-word stage (combinatorial signalling), and eventually to (compositional) language.[9] However, as was mentioned in the introduction, prioritising linguistic compositionality as an explanatory target gives rise to significant theoretical and practical problems.

The novel approach to the evolution of language suggested here prioritises reflexivity as an explanatory target. On this account, simple communicative capacities evolve alongside cognitive capacities. Signals may become functionally referential, referring to concrete objects in the world. Once individuals are able to make use of proto-concepts, they can refer to *abstracta*. Therefore, they can refer to communicative contexts, giving rise to proto-reflexivity. This ability means that they can

[8] See empirical work in Wynn (1992); Simon et al. (1995); Koechlin et al. (1998); de Walle et al. (2001); Feigenson et al. (2002). See Wynn (1998); Spelke (1998) for reviews of this literature.
[9] See, e.g., Bickerton (1990); Jackendoff (1999); Progovac (2015).

influence future communicative behaviour via communication. Such capacities may evolve by modular composition and related processes. Furthermore, it has been demonstrated that reflexivity gives rise to functional composition (compositional syntax) as a byproduct of these processes (LaCroix, 2019b).

Several recent works in the signalling game literature have demonstrated that modular compositional processes, like the ones described here, are more efficient and more effective for evolving or learning communication conventions than learning novel dispositions from scratch, often by orders of magnitude (Barrett, 2016, 2017, 2020; Barrett and Skyrms, 2017; LaCroix, 2019b, 2020c; Barrett et al., 2020).

Furthermore, reflexivity does not succumb to the same problems that compositionality does, as an explanatory target. It was mentioned in the introduction that compositionality, as it is discussed in the literature, fails to maintain sensitivity to role-asymmetries between producers and interpreters of signals (LaCroix, 2020a); however, for reflexivity, this role-asymmetry is built-in via functional reference (Definitions 2.1; 2.2), which accounts for these differences by definition. Furthermore, there are no empirical precursors to compositionality (LaCroix, 2019a), whereas the processes by which reflexivity evolves are supported by significant empirical evidence. Finally, compositionality is a binary property of language (Berwick and Chomsky, 2011), meaning that there is no gradualist explanation of the evolution of compositionality; in contrast, both reflexivity and the processes by which it might arise are graded notions. In non-reflexive functionally-referential systems, signals refer to states; in proto-reflexive functionally-referential systems, signals refer to communicative contexts; and in reflexive language, words refer to linguistic entities. So, reflexivity is graded, but the processes by which it arises are also graded—appropriation is simpler than analogical reasoning, which is simpler than modular composition.

Finally, compositionality is focused too internally on language and syntax itself, so explanations do not (or at least need not) take account of related cognitive and social mechanisms that are important factors in the evolution of language. On the other hand, reflexivity does. Therefore, there are significant practical and theoretical reasons to replace compositionality with reflexivity as an explanatory target for language origins research.

References

Barrett, H. Clark and Robert Kurzban (2006). Modularity in Cognition: Framing the Debate. *Psychological Review*, 113(3): 628–647.

Barrett, Jeffrey (2016). On the Evolution of Truth. *Erkenntnis*, 81: 1323–1332.

Barrett, Jeffrey (2017). Truth and Probability in Evolutionary Games. *Journal of Experimental and Theoretical Artificial Intelligence*, 29(1): 219–225.

Barrett, Jeffrey A. (2020). Self-assembling games and the evolution of salience. *British Journal for the Philosophy of Science*. Forthcoming.

Barrett, Jeffrey A. and Brian Skyrms (2017). Self-Assembling Games. *The British Journal for the Philosophy of Science*, 68(2): 329–353.

Barrett, Jeffrey A., Brian Skyrms, and Calvin Cochran (2020). Hierarchical Models for the Evolution of Compositional Language. *Philosophy of Science*. Forthcoming.

Berwick, Robert C. and Noam Chomsky (2011). The Biolinguistic Program: The Current State of its Development. In Sciullo, A. M. Di and C. Boeckx, editors, *The Biolinguistic Enterprise: New Perspectives on the Evolution and Nature of the Human Language Faculty*, pages 19–41. Oxford University Press, Oxford.

Bickerton, Derek (1990). *Language and Species*. University of Chicago Press, Chicago.

Blough, Donald S. (1959). Delayed Matching in the Pigeon. *Journal of the Experimental Analysis of Behavior*, 2(2): 151–160.

Bonte, Élodie, Caralyn Kemp, and Joël Fagot (2014). Age Effects on Transfer Index Performance and Executive Control in Baboons (*Papio papio*). *Frontiers in Psychology*, 5: 188.

Cangelosi, Angelo and Domenico Parisi (2002). Computer Simulation: A New Scientific Approach to the Study of Language Evolution. In *Simulating the Evolution of Language*, pages 3–28. Springer, London.

Carruthers, Peter (2002). The Cognitive Functions of Language. *Behavioral and Brain Sciences*, 25(3): 657–725.

Coltheart, Max (1999). Modularity and Cognition. *Trends in Cognitive Sciences*, 3(3): 115–120.

Dawkins, Richard and John R. Krebs (1978). Animal signals: Information or Manipulation? In Krebs, J. R. and N. B. Davies, editors, *Behavioural Ecology*, pages 282–309. Blackwell Scientific Publications, Oxford.

de Walle, Gretchen A. Van, Susan Carey, and Meredith Prevor (2001). Bases for Object Individuation in Infancy: Evidence from Manual Search. *Journal of Cognition and Development*, 1(3): 249–280.

Ellis, Henry Carlton (1965). *The Transfer of Learning*. The Macmillan Company, New York.

Fagot, Jöel, Edward A. Wasserman, and Michael E. Young (2001). Discriminating the Relation Between Relations: The Role of Entropy in Abstract Conceptualization by Baboons (*Papio papio*) and Humans (*Homo sapiens*). *Journal of Experimental Psychology: Animal Behavior Processes*, 27(4): 316–328.

Feigenson, Lisa, Susan Carey, and Elizabeth S. Spelke (2002). Infants' Discrimination of Number vs. Continuous Extent. *Cognitive Psychology*, 44(1): 33–36.

Ferster, Charles Bohris (1960). Intermittent Reinforcement of Matching to Sample in the Pigeon. *Journal of the Experimental Analysis of Behavior*, 3(3): 259–272.

Fitch, W. Tecumseh (2017). Empirical Approaches to the Study of Language Evolution. *Psychonomic Bulletin & Review*, 24(1): 3–33.

Flemming, Timothy M., Roger K. R. Thompson, Michael J. Beran, and David A. Washburn (2011). Analogical Reasoning and the Differential Outcome Effect: Transitory Bridging of the Conceptual Gap for Rhesus Monkeys (*Macaca mulatta*). *Journal of Experimental Psychology: Animal Behavior Processes*, 37(3): 353–360.

Fodor, Jerry (1983). *The Modularity of Mind*. MIT Press, Cambridge MA.

Fodor, Jerry (1984). Observation Reconsidered. *Philosophy of Science*, 51: 23–43.

Fodor, Jerry (2000). *The Mind Doesn't Work That Way*. MIT Press, Cambridge MA.

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 298: 1569–1579.

Hockett, Charles F. (1960). The Origin of Speech. *Scientific American*, 203: 88–111.

Hung, Woei (2013). Problem-Based Learning: A Learning Environment for Enhancing Learning Transfer. *New Directions for Adult and Continuing Education*, 137: 27–38.

Hurford, James R. (2007). *Language in the Light of Evolution I: The Origins of Meaning*. Oxford University Press, Oxford.

Jackendoff, Ray S. (1999). Possible Stages in the Evolution of the Language Capacity. *Trends in Cognitive Sciences*, 3: 272–279.

Katz, J. S., A. A. Wright, and J. Bachevalier (2002). Mechanisms of Same/Different Abstract-Concept Learning by Rhesus Monkeys (*Macaca mulatta*). *Journal of Experimental Psychology: Animal Behavior Processes*, 28(4): 358–368.

Kight, Caitlin R., John M. McNamara, David W. Stephens, and Sasha R. X. Dall (2013). Communication as information use: Insights from statistical decision theory. In Stegmann, Ulrich E., editor, *Animal Communication Theory: Information and Influence*, pages 89–112. Cambridge University Press, Cambridge.

Koechlin, Etienne, Stanislas Dehaene, and Jacques Mehler (1998). Numerical Transformations in Five-Month-Old Human Infants. *Mathematical Cognition*, 3(2): 89–104.

LaCroix, Travis (2019a). Biology and compositionality: Empirical considerations for emergent-communication protocols.

LaCroix, Travis (2019b). Using logic to evolve more logic: Composing logical operators via self-assembly. *British Journal for the Philosophy of Science*. Forthcoming.

LaCroix, Travis (2020a). Accounting for polysemy and role asymmetry in the evolution of compositional signals. unpublished manuscript.

LaCroix, Travis (2020b). *Complex Signals: Reflexivity, Hierarchical Structure, and Modular Composition*. PhD dissertation, University of California, Irvine.

LaCroix, Travis (2020c). The correction game or, how pre-evolved communicative dispositions might affect communicative dispositions. unpublished manuscript.

Lewis, David (2002/1969). *Convention: A Philosophical Study*. Blackwell, Oxford.

Macedonia, Joseph M. and Christopher S. Evans (1993). Variation Among Mammalian Alarm Call Systems and the Problem of Meaning in Animal Signals. *Ethology*, 93: 177–197.

Martinez, Pedro (2018). The Comparative Method in Biology and the Essentialist Trap. *Frontiers in Ecology and Evolution*, 6(130): 1–5.

Progovac, Ljiljana (2015). *Evolutionary Syntax*. Oxford University Press, Oxford.

Pugh, Kevin J. and David A. Bergin (2006). Motivational Influences on Transfer. *Educational Psychologist*, 41(3): 147–160.

Resnik, David B. (1991). How-Possibly Explanations in Biology. *Acta Biotheoretica*, 39(2): 141–149.

Robbins, Philip (2017). Modularity of mind. In Zalta, Edward N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition.

Rumbaugh, Duane M. (1970). Learning Skills of Anthropoids. In Rosenblum, L., editor, *Primate Behavior: Developments in Field and Laboratory Research*, pages 2–70. Aldine, New York.

Rumbaugh, Duane M. (1971). Evidence of Qualitative Differences in Learning Processes Among Primates. *Journal of Comparative and Physiological Psychology*, 76(2): 250–255.

Rumbaugh, Duane M. (1995). Primate Language and Cognition: Common Ground. *Social Research*, 62(3): 711–730.

Rumbaugh, Duane M. and James L. Pate (1984a). Primates' Learning by Levels. In Greenberg, G. and E. Tobach, editors, *Behavioral Evolution and Integrative Levels*, pages 221–240. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rumbaugh, Duane M. and James L. Pate (1984b). The Evolution of Cognition in Primates: A Comparative Perspective. In Roitblat, H., T. G. Bever, and H. S. Terrace, editors, *Animal Cognition*, pages 569–587. Lawrence Erlbaum Associates, Hillsdale, NJ.

Scarantino, Andrea (2013). Animal Communication as Information-Mediated Influence. In Stegmann, Ulrich E., editor, *Animal Communication Theory: Information and Influence*, pages 63–87. Cambridge University Press, Cambridge.

Schunk, Dale H. (2004). *Learning Theories: An Educational Perspective*. Pearson, Upper Saddle River, NJ, 4 edition.

Seyfarth, Robert M., Dorothy L. Cheney, and Peter Marler (1980). Vervet Monkey Alarm Calls: Semantic Communication in a Free-Ranging Primate. *Animal Behaviour*, 28(4): 1070–1094.

Sherry, David F. and Daniel L. Schacter (1987). The Evolution of Multiple Memory Systems. *Psychological Review*, 94(4): 439–454.

Shettleworth, Sara J. (2012). Modularity, Comparative Cognition and Human Uniqueness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1603): 2794–2802.

Sievers, C. and T. Gruber (2016). Reference in human and non-human primate communication: What does it take to refer? *Animal Cognition*, 19(4): 759–768.

Simon, Tony J., Susan J. Hespos, and Philippe Rochat (1995). Do Infants Understand Simple Arithmetic? A Replication of Wynn (1992). *Cognitive Development*, 10(2): 253–269.

Skinner, Burrhus F. (1950). Are Theories of Learning Necessary? *Psychological Review*, 57: 193–216.

Skyrms, Brian (2010a). *Signals: Evolution, Learning, & Information*. Oxford University Press, Oxford.

Skyrms, Brian (2010b). The Flow of Information in Signaling Games. *Philosophical Studies*, 147(1): 155–165.

Spelke, Elizabeth S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, 21(2): 181–200.

Spelke, Elizabeth S. (2003). What Makes Us Smart? Core Knowledge and Natural Language. In Gentner, Dedre and Susan Goldin-Meadow, editors, *Language in Mind: Advances in the Investigation of Language and Thought*, pages 277–311. MIT Press, Cambridge, MA.

Sperber, Dan (1994). The Modularity of Thought and the Epidemiology of Representations. In Hirschfeld, L. A. and S. A. Gelman, editors, *Mapping the Mind*, pages 39–67. Cambridge University Press, Cambridge, MA.

Sperber, Dan (2002). In Defense of Massive Modularity. In Dupoux, I., editor, *Language, Brain, and Cognitive Development*, pages 47–57. MIT Press, Cambridge, MA.

Stegmann, Ulrich E. (2013). *Animal Communication Theory: Information and Influence*. Cambridge University Press, Cambridge.

Thompson, R. K. R. and D. L. Oden (2000). Categorical Perception and Conceptual Judgments by Non-Human Primates: the Paleological Monkey and the Analogical Ape. *Cognitive Science: A Multidisciplinary Journal*, 24(3): 363–396.

Wasserman, Edward A., Michael E. Young, and Jöel Fagot (2001). Effects of Number of Items on the Baboon's Discrimination of Same from Different Visual Displays. *Animal Cognition*, 4(3-4): 163–170.

Wynn, Karen (1992). Addition and Subtraction by Human Infants. *Nature*, 358: 749–750.

Wynn, Karen (1998). Psychological Foundations of Number: Numerical Competence in Human Infants. *Trends in Cognitive Sciences*, 2(8): 296–303.

Zuberbühler, Klaus, Dorothy L. Cheney, and Robert M. Seyfarth (1999). Conceptual Semantics in a Non-Human Primate. *Journal of Comparative Psychology*, 113(1): 33–42.

The Evidence-Observation Distinction in Observation Selection Effects

Word Count: 4955

Abstract

Previous discussions of observation selection effects (OSEs) have ignored the distinction between observation and evidence. Evidence for a hypothesis, I argue, is distinct from the observation of that evidence. This shows that the fact that evidence is unobservable does not entail that the evidence does not obtain. What is required for an OSE is that evidence is guaranteed, not that counter-evidence is unobservable. With the evidence-observation distinction in hand, apparent counterexamples fail. I then show that observer perspective can change whether or not an agent is subject to an OSE, even when knowledge is shared between perspectives.

1 Introduction

In this paper I defend the entailment model of observation selection effects (OSEs). This simple model states that when background conditions, in conjunction with the hypotheses under consideration, entail evidence E, then E does not favor either of the

hypotheses.[1] However, this model for understanding how OSEs work has been subject to several apparent counterexamples, including FIRING SQUAD and a modified version of Eddington's Fish example.

What these, and other discussants, have presupposed, however, is that evidence and observation of evidence are equivalent. I argue that evidence and observation of evidence are importantly distinct. Evidence is a fact or state of affairs, about the world, that obtains. But the fact that a state of affairs obtains does not guarantee that it is observed. In many cases evidence obtains, but we do not, or cannot, observe it. With this distinction clarified, the entailment model is shown to handle the apparent counterexamples deployed against it.

I then present an example from Francis Bacon which shows the surprising result that observer perspective can change whether or not an agent is subject to an OSE, even when both perspectives share all the same knowledge.

2 Eddington's Fish and the Entailment Model

Discussions of observation selection effects (OSEs) rightly begin with an example adapted from Eddington's 1939 *Philosophy of Physical Science*. (Eddington 1939)[2] In this example, FISHING, we imagine a biologist attempting to distinguish between two

[1]It is not necessary that selection biases, more generally, require that the background and hypotheses *entail* the evidence. For a more general discussion of the phenomena of selection bias see Berkson's Paradox.

[2]This version due to Sober 2003, 41f.

hypotheses:

L: All fish in this pond are longer than 10 inches.

S: Half of the fish in this pond are longer than 10 inches, the other half are shorter than 10 inches.

The biologist then observes this evidence:

E: All fish caught in this pond were longer than 10 inches.

How should we evaluate the hypotheses in light of E? One way would be to express the relationship between the evidence and hypotheses would be through the Law of Likelihood:

Law of Likelihood: E is evidence for H_i over H_j iff $pr(E|H_i) > pr(E|H_j)$

In this case, since all the fish caught were larger than 10", and this is more likely if L is true rather than S, we get the following inequality:

$$pr(E|L) > pr(E|S)$$

Thus, by the law of likelihood, E is evidence for L over S.

But now a further fact about how the observations were made is revealed:

N: The net used to catch fish in this pond always catches fish, if it can, but it can't catch fish smaller than 10 inches because of the size of the holes in the net.[3]

[3]Most authors have given N as, "The net used can't catch fish smaller than 10 inches

Now E no longer appears to be evidence for L over S because:

$$pr(E|L \wedge N) = pr(E|S \wedge N) = 1$$

The simplest explanation for what has happened is that,

$$N \wedge L \vDash E \text{ and } N \wedge S \vDash E$$

Since N, in conjunction with either L or S, ensures that any fish caught will be larger than 10", it entails E. The alternative, $\neg E$, that some fish caught in this pond were shorter than 10", is ruled out.

In general:

> If, for background conditions B and hypotheses H_1 and H_2, $B \wedge H_1 \vDash E$ and $B \wedge H_2 \vDash E$, then E is not evidence for H_1 over H_2.

Since it is necessary to consider background conditions, we must supplement the Law of Likelihood with a total evidence requirement which makes it explicit that we must take these conditions into account:

> Law of Likelihood*: E is evidence for H_1 over H_2 with respect to background conditions B iff $pr(E|H_1 \wedge B) > pr(E|H_2 \wedge B)$

This model of observation selection effects is the one I clarify and defend in the rest of the paper. In the next section, I explain the FIRING SQUAD counterexample to the

---

because of the size of the holes in the net," but then need to build in a "catch" condition so that $N \vDash E$.

entailment model. I then show that, once we distinguish evidence from observation, the entailment model gives the right result for Firing Squad after all.

3 Evidence, Observation, and Firing Squads

3.1 Firing Squad

The first objection to the simple model above is Firing Squad (Sober 2003, 44f.). In this example, imagine a prisoner faced with a firing squad. Two hypotheses are being considered:

Aim: The firing squad is aiming at the prisoner.

Avoid: The firing squad is aiming to avoid hitting the prisoner.

After the shots are fired and the smoke has cleared, the prisoner makes her observation:

Alive: The prisoner is alive.

Since it is unlikely that the prisoner would still be alive, were the executioners aiming, but quite likely that she would be alive, if they were avoiding her, it seems that:

$$pr(\text{Alive}|\text{Avoid}) > pr(\text{Alive}|\text{Aim})$$

Thus her survival seems to be evidence that the executioners were aiming to miss.

But, if the entailment model formulated above is correct, there appears to be an OSE at work in Firing Squad. The way the evidence was gathered guarantees that, if the prisoner observes anything, she is guaranteed to observe that she is alive. The background necessary for the prisoner to observe that she is alive is:

Survive: The prisoner survives.

Survive clearly entails Alive, so:

$$pr(\text{Alive}|\text{Aim} \wedge \text{Survive}) = pr(\text{Alive}|\text{Avoid} \wedge \text{Survive}) = 1$$

But this will not do. It seems clear that the prisoner would be correct in taking her survival to be evidence that the guards aimed to miss. What has gone wrong?

3.2 Evidence and Observation

In order to understand what has gone wrong, we must give an account of the difference between observation and evidence. These two concepts have frequently been conflated. From an internalist perspective it is often assumed that observations are the only things which could possibly be *used as* evidence. Evidence, after all, must be accessible to the agent and therefore must be a mental state (or similar). The only candidate for evidence about the world, then, is our phenomenal experience of it—namely observations.

But there is more to the story. If we take observations to be evidence then we can, classically, only have two evidential states with respect to E: either we have the evidence E or we do not. But once we move outside of the agent and into an external world, rich in evidence, that is not how observations *of* evidence work. No doubt we *do* need to observe in order to incorporate evidence, but that is not what evidence is.

Evidence is, on my account, a state-of affairs or fact about the world. It can be, and often is, independent of our knowledge, awareness, or observation of it. As a state-of-affairs it either is, or is not, the case. Thus for any state-of-affairs E, either E or $\neg E$.

Observation of that E, on the other hand, is distinct from E. While E or $\neg E$ is necessary, our observation of E or $\neg E$ is not. We might observe E, observe $\neg E$, or we might fail to observe anything, with respect to E.

The test of whether some claim is an observation claim or an evidence claim will, then, as a first pass, be whether it obeys excluded middle. If the claim about φ does obey excluded middle, then it is an evidence claim. If it does not and there is a third option—failure to observe φ—then it cannot be a claim about evidence and must be a claim about observation.

Here's an example to help understand this distinction: An astrobiologist seeks evidence of life on other planets, such as O_2 concentrations in the atmosphere. There is a fact-of-the-matter about the O_2 concentration. It is there whether anyone ever knows about it. This evidence, however is difficult to observe. It may never be observed. But this does not mean that a high O_2 concentration is not evidence for life. It merely means that our observation of this evidence is contingent.

As a first pass at giving a logic of observation, let me propose the observation operator '$\bigcirc$'.[4] $\bigcirc_\alpha\varphi$ should be read as the tenseless claim that α observes φ. One might plausibly know $\bigcirc_\alpha\varphi$ without being the agent α and without observing φ oneself. Crucially, $\bigcirc$ will have the following properties:

1. $\bigcirc_\alpha\varphi \vDash \varphi$

2. $\neg \bigcirc_\alpha \varphi \nvDash \neg\varphi$

[4] $\bigcirc$ may be understood as similar to the epistemic modal operator K and will have similar properties, though developing the connections will require further work.

Thus observation is factive; observation of some evidence E, by an agent, α, entails E (or alternatively, entails that E is the case). But failure to observe E does not entail that there is no evidence E (or alternatively, that E fails to obtain). This distinction is, I think, easily understood, but easily overlooked.

The same distinction between observation and evidence may be elucidated in terms of conditionalization upon learning new evidence:

> Conditionalization: for any time t_i and later time t_j, if proposition E represents everything the agent learns between t_i and t_j and $pr_i(E) > 0$, then for any H, $pr_j(H) = pr_i(H|E)$ (Titelbaum 2015, 92).[5]

Note that conditionalization is put in terms of *learning* E, not in terms of whether E or $\neg E$. Just as with observation, failing to learn E between t_i and t_j does not entail $\neg E$. One does not update upon failing to learn anything.[6] Thus failing to observe E is equivalent to learning nothing between t_i and t_j.

Conditionalization tells us that Bayes' rule is about what one would do *if* one had the evidence, not simply what one believes already. It sets a norm that, if one had some evidence, one ought do such-and-such with that evidence. It does not guarantee that one has that evidence. This diachronic norm forces us to consider the act of learning—observing—new evidence. Rather than just considering evidence as a static phenomenon, the observation of evidence causes a change in our epistemic state. Thus

[5]Other updating norms are, of course, possible, such as Jeffrey conditionalization.

[6]Note, however, that one might learn, between t_i and t_j, that one has (or has not) made an observation. φ may be a complex statement which contains $\bigcirc$.

we should be considering two states: the state prior to receiving the evidence, which is when the law of likelihood tells us how we *should* react to the evidence, and the state after receiving the evidence, which is when we *do* incorporate that evidence.

Since Bayesians posit this tight link between synchronic conditional credences and posterior diachronic credences, how one ought to update on evidence is built in from the beginning. One can, therefore, evaluate how one ought to react upon learning E without having to observe E. Similarly, one retains this judgment about how one should evaluate E whether or not one does, or even can, observe E. Thus the evidence and observation distinction is built into the Bayesian approach.

This parallel between evidence-observation and conditionalization will be a useful heuristic going forward.

3.3 Resolving Firing Squad

Let's take a closer look at Firing Squad to see how the observation-evidence distinction is relevant.

First, let us use conditionalization to review the situation. Recall that each of Aim and Avoid, in conjunction with Survive entail Alive, thus Alive seemed to provide no evidence for Avoid over Aim. But as we are considering how evidence is used to update beliefs, we should consider the prisoner's epistemic situation before and after the shots are fired.

It may be easiest to consider the perspective of a bystander to Firing Squad. This bystander will not be subject to an OSE. Her life and future observations are not threatened by the executioners' guns. Thus he can legitimately have this likelihood

argument in mind at t_0:

$$pr(\text{Alive}|\text{Avoid}) > pr(\text{Alive}|\text{Aim})$$

Thus, upon learning, between t_0 and t_1, whether or not the prisoner is alive, he can update on that evidence. Both Alive and ¬Alive are possible states of affairs that come into being between t_0 and t_1. He learns something new when the smoke clears and the prisoner has survived.

The prisoner is in exactly the same evidential position at t_0, prior to the volley.[7] Just as the bystander is in suspense as to whether Alive or ¬Alive, the prisoner too is in suspense. She does not know what will happen. She knows that she will not observe that she is not alive, but this does not rule out the possibility that she does not survive. Thus, when the smoke clears, she learns something new and surprising—she is alive! She then updates on the information learned between t_0 and t_1, increases her credence in Avoid, and lowers her credence in Aim.

Let us put this in terms of observation and evidence. What will the prisoner's background conditions include? Survive (the prisoner survives) will not be in the prisoner's background conditions. Survive would presuppose that the only evidence possible is Alive. But this is not the case. Alive and ¬Alive are both possible states of affairs. What will be among her background conditions is Survive*:

[7]Contra Sober who argues that, because the prisoner cannot make the observation that she is not alive, the prisoner and bystander are in different evidential positions (Sober 2003, p. 46, 50n20f.).

Survive*: I will not observe that I am not alive.[8]

In symbols, for prisoner p:

$$\neg \bigcirc_p \neg \text{Alive}$$

This states that the prisoner, p, will not observe that she is not alive. Thus the correct likelihood argument will be:

$$pr(\text{Alive}|\text{Avoid} \wedge \neg \bigcirc_p \neg \text{Alive}) > pr(\text{Alive}|\text{Aim} \wedge \neg \bigcirc_p \neg \text{Alive})$$

Recall that $\neg \bigcirc \text{Alive} \not\models \neg \text{Alive}$. Survive* does not entail Alive. Survive* thus makes no difference to the likelihood argument. The fact that the prisoner will not *observe* that she is alive does not entail that she will not *be* alive. Thus, since it is not guaranteed that she survives, her survival can count as evidence that the soldiers were not aiming to hit her.

Once we recognize the role that the evidence-observation distinction is making, the entailment model gives the correct result for Firing Squad after all.

3.4 The No-Observation Objection

But this response leads to another objection to the entailment model. If we are correct about how the entailment model ought to respond to Firing Squad, then it seems as though our intrepid ichthyologist is also not guaranteed to observe that there are large fish in the net.

[8]Survive* is equivalent to Weisberg's S': If I observe whether I survive, I will observe that I survive (Weisberg 2005, 816).

The objection is this: it is possible that the biologist dies, or simply fails to return to the pond to look in her net, thus the evidence is no more guaranteed for her than for the hapless prisoner above. The biologist knows that she will not observe anything other than large fish, but this does not entail that she will observe large fish. In the same way, the prisoner knows she will not observe that she is not alive, but this does not entail that she will be alive. If this is right then the biologist's background conditions do not seem to entail E (Weisberg 2005, 817). Since the evidence is not entailed, there will be no OSE and the fish in the net *are* evidence for L over S.

In order to answer this we must be careful to specify what the evidence is in FISHING and what is in the biologist's background conditions. Recall what N says:

> N: The net used to catch fish in this pond always catches fish, if it can, but it can't catch fish smaller than 10 inches because of the size of the holes in the net.[9]

And this, in conjunction with either L or S, does entail E, which states:

> E: All fish caught in this pond were longer than 10 inches.

Thus the biologist knows, given N, without needing to observe the evidence, what the evidence is: there are large fish in the net. N therefore entails, not that she observes large fish ($\bigcirc E$), but the evidence (E) itself: All fish caught in the pond are larger than 10". The biologist will, of course, also know that she will not observe anything other than large fish ($\neg \bigcirc \neg E$), but this is irrelevant as she knows, without observing E, that E obtains.

[9]See footnote 3.

Conditionalization can similarly explain the problem. If the biologist dies, or otherwise fails to return to the net, she learns neither E nor $\neg E$. Thus she cannot change her credences $pr(L)$ or $pr(S)$ on the basis of E or $\neg E$. One cannot update without observing evidence!

The No-Observation objection seems to depend on one of two confusions. First, we might think that, in order to generate an OSE, $\bigcirc E$ must have been in the biologist's background conditions. Second, we might have thought that $\neg \bigcirc \neg E \vDash E$. But both of these assumptions are incorrect and both depend on a conflation of evidence and observation.

Sober seems to make this first mistake, stating: "If you fish with Eddington's net, you are guaranteed to observe that the net contains fish that all are over 10 inches long" (Sober 2009, 77). If we mistake observation for evidence then it is easy to assume that N is equivalent to $\bigcirc E$. That is, the fact that the net will contain large fish is the same as the claim that we will observe that the net contains large fish. But of course, as the no-observation objection shows, the biologist is not guaranteed to observe anything! The net, however, is guaranteed to contain fish over 10".

Second, we might have thought that $\neg \bigcirc \neg E \vDash E$, and since $\neg \bigcirc \neg E$ is in the biologist's background, that is the reason E is entailed. But this is, once again, to think that failure to observe small fish is an observation of large fish.

The lesson here is that, if we were already inclined to think that evidence must be observed in order to make any difference to our arguments, we will think that there is no harm in using 'evidence' and 'observation' interchangeably. Not observing E will sound a lot like observing $\neg E$. Similarly, we might argue that observation entails evidence, thus there is no harm in using them interchangeably. But observation statements are not

equivalent to evidence statements and thus there is harm in conflating evidence and observation. We will think that necessitated evidence necessitates observation or that failure to observe is itself an observation. This is a mistake. What is the case in FISHING is that the *observation* that there are large fish in the net is not guaranteed, but the *evidence*—the fact that there will be fish in the net—is guaranteed. The fact that evidence can be guaranteed—and we can know that it is guaranteed—makes it such that evidence can make a difference to OSEs, even without being observed.

To sum up, in FISHING the evidence is guaranteed, but the observation of it is not. In FIRING SQUAD neither the evidence nor the observation is guaranteed. The objection was that observation is not guaranteed, and this is true. But I've argued that it is the *evidence* that must be guaranteed by background conditions, not the observation of it. Thus the biologist in FISHING is subject to an OSE, while the prisoner in FIRING SQUAD is not.

4 Survivor Bias and the Power of Prayer

The evidence-observation distinction allows us to see a further consequence of observation selection-effects: parties can differ in what their evidence supports, even if both parties know all the same facts. To illustrate this, let's start with a case of shipwreck survivors and the power of prayer given by Sir Francis Bacon in *Novum Organum*:

> It was well answered by him who was shown in a temple the votive tablets suspended by such as had escaped the peril of shipwreck, and was pressed as to whether he would then recognize the power of the gods, by an inquiry, But where are the portraits of those who have perished in spite of their vows?

(Bacon 1620/2000, XLVI)

There is an OSE at work here. Clearly, only survivors give testimony (in the form of votive tablets). Those whose prayers were not answered did not survive to give testimony. Thus it is guaranteed that the evidence of the votives will always be from sailors who prayed and survived. Bacon's story is well explained by the entailment model of OSEs. A background condition for the testimony is the survival of the testifier. Thus the background condition entails that a sailor who leaves a votive must have survived, no matter the efficacy of prayer.

Now, to see the surprising result, let's consider a hypothetical conversation between a single sailor and the skeptical temple visitor, call it SHIPWRECK:

> Sailor: As the ship was sinking I wasn't sure I was going to make it. So I prayed and, lo and behold, I was saved! Surely, as I might not have been saved, my survival provides evidence that my prayer was effective.
>
> Visitor: I'm very happy you survived, but I'm sorry to say that your survival is no evidence that your prayer was effective. After all, if you hadn't survived, you wouldn't be standing here telling me your story. You couldn't have told me that you prayed and weren't saved, thus, because the fact that I met you entails that you survived, it does not tell me anything about whether prayer is effective.
>
> Sailor: But you must admit that *I* might not have been saved, and that's all that the argument requires. It's not the fact that we met that matters, but that I survived when I might not have.
>
> Visitor: True, the fact that we met doesn't matter to *your* argument. But

> that fact matters very much to my argument. You must admit that I wouldn't know anything about your prayer or your shipwreck without your being able to tell me. I am not surprised that, given I met someone who was in a shipwreck, I met one of the survivors. I could not possibly have met someone who did not survive.
>
> Sailor: Then it seems we're at an impasse. I can tell you everything about my situation and it doesn't change the fact that you could only have learned these things by meeting a shipwreck survivor. You can tell me all about the fact that no other evidence was possible for you and it changes nothing for me. How strange that we agree on all the facts, but differ in what our evidence supports.

What has happened here? Unlike the prisoner and bystander in Firing Squad, in which neither of the parties was subject to an OSE, in Shipwreck the sailor is not subject to an OSE, while the visitor is. Let's take a closer look at the sailor and visitor's likelihood arguments:

Let our two hypotheses be:

> Effective: Prayer is efficacious.
>
> ¬Effective: Prayer is inefficacious.

And the sailor's evidence:

> Alive: Sailor s is alive.

Formally, we may now give the sailor's argument as:

$$pr(\text{Alive}|\text{Effective} \wedge \neg \bigcirc_s \neg\text{Alive}) > pr(\text{Alive}|\neg\text{Effective} \wedge \neg \bigcirc_s \neg\text{Alive})$$

The sailor could not have observed that she did not survive, but as this does not entail that she does survive, it is no challenge to the argument, according to the entailment model. She is in an analogous position to the prisoner in FIRING SQUAD.

The visitor's situation is crucially different in two ways. First, there is an additional relevant fact in her background conditions:

> Meet: The visitor, v, meets a sailor, σ, who has been shipwrecked, but v cannot meet sailors who did not survive shipwreck.

This is parallel to the fact N in FISHING. Just as N guarantees that there will be fish in the net, the visitor does meet a sailor who has been shipwrecked. And just as N guarantees that the fish caught will be larger than 10", the visitor can only meet sailors who have survived shipwreck.

But note that, while Meet does guarantee that the visitor meets a sailor, it does not guarantee that he meets this particular sailor, s. The claims in N and Meet *describe* the evidence, but do not mention the particular fish that was caught or the particular sailor who was met. This leads us to the second difference between the sailor and the visitor's situations: the visitor does not care which particular sailor he meets. In the visitor's case, then, the evidence is:

> Alive*: The sailor, σ, that v meets, is alive.

The evidence, for the visitor, is not that some particular sailor, s, survived, but that a

sailor, σ, survived and is met. Similarly, in FISHING, no particular fish was guaranteed to be caught, but the fish that was caught was guaranteed to be large. We now have all the necessary ingredients to give the visitor v's argument:

$$pr(\text{Alive*}|\text{Effective} \wedge \text{Meet}) = pr(\text{Alive*}|\neg\text{Effective} \wedge \text{Meet}) = 1$$

The two sides of the argument are equal because Meet $\vDash$ Alive*. Thus there is an OSE for the visitor.

Note that this is different from the bystander's position in FIRING SQUAD. In that case the evidence was the same for both prisoner and bystander. Prisoner P's survival was what mattered to both the prisoner and bystander in that case. Thus the evidence that the prisoner survived was not guaranteed.

The fact that the visitor is subject to an OSE, while the sailor is not, remains true when the two meet and share information. It is not relevant to the sailor that Meet be known, because Meet only entails that v meet some sailor σ, not that she meet sailor s. Similarly, the fact that sailor s survived is crucial to the sailor's argument—if she had not survived, Effective would not be supported. But the fact that sailor s survived is irrelevant to the visitor's argument—if sailor s had not survived, the visitor would have met some other sailor who had survived. Thus while the two share all the same facts, they do not take those facts to provide equal support for the efficacy of prayer.

This surprising result follows because the arguments that the sailor and visitor make are different and use different statements of the evidence to come to their respective conclusions. The sailor's argument uses the *de re* claim that she survives, while the vistor's argument uses the *de dicto* claim that someone survives. More needs to be said

about whether this can result from a difference in the facts known by the agents. It is at least plausible that the difference rests in a difference in known facts such as, "I am subject to an OSE," said by v, which differs in content from "v is subject to an OSE." Such facts will be unsharable by the agents.

5 Conclusions

In this paper I have distinguished observation from evidence and shown how this can help clarify when there is, or is not, an OSE. I have given three kinds of scenarios:

1. Firing Squad Bystander: Can observe E and can observe $\neg E$. No OSE.
2. Firing Squad Prisoner and Shipwreck Sailor: Cannot observe $\neg E$, but E is not guaranteed. No OSE.
3. Shipwreck Visitor and Fishing: E is guaranteed. OSE.

These three possibilities make it clear that the unobservability of evidence is necessary, but not sufficient for an OSE.

Although I suggest the development of a modal treatment of observation, there is a quick heuristic for distinguishing an evidence claim from observation claim. If the claim about φ does obey excluded middle, then it is an evidence claim. If it does not and there is a third option—failure to observe φ—then it cannot be a claim about evidence and must be a claim about observation.

Finally, observer perspective seems to matter in cases such as the two Shipwreck scenarios. Whether this results from an unshared—or unsharable—fact or results from the accessibility of arguments to agents is an avenue for further exploration.

References

Bacon, Francis. 1620/2000. *The New Organon.* Edited by Lisa Jardine and Michael Silverthorne.

Eddington, Arthur Stanley. 1939. *The Philosophy of Physical Science.* CUP Archive.

Sober, Elliott. 2003. God and Design: The teleological argument and modern science. In *God and design: the teleological argument and modern science,* edited by Neil A Manson. Routledge.

———. 2009. Absence of Evidence and Evidence of Absence: Evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies* 143 (1): 63–90.

Titelbaum, Michael. 2015. Fundamentals of Bayesian Epistemology. *Unpublished manuscript.*

Weisberg, Jonathan. 2005. Firing Squads and Fine-Tuning: Sober on the design argument. *The British Journal for the Philosophy of Science* 56 (4): 809–821.

Infrared Cancellation and Measurement

Michael E. Miller[†]

Quantum field theories containing massless particles such as photons and gluons are divergent not just in the ultraviolet, but also in the infrared. Infrared divergences are typically regarded as less conceptually problematic than ultraviolet divergences because there is a reasonably straightforward cancellation mechanism that renders measurable physical observables such as decay rates and cross-sections infrared finite. In this paper, I scrutinize the restriction to *measurable* physical observables that is required to make the cancellation mechanism applicable. I argue that this restriction does not necessitate a retreat to operationalism about the meaning of the theory as one might reasonably have worried, but it does call attention to a collection of underappreciated conceptual issues lurking in the infrared regime of quantum field theories with massless particles.

1. Introduction. The structural core of non-relativistic quantum mechanics is reasonably well agreed upon. It includes states defined on a Hilbert space, operators on that space to represent observables, the Schrödinger dynamics, and the Born rule for determining probabilities for the outcomes of experiments.[1] This structural core provides an algorithm for extracting empirical predictions from the theory. Interpretive debates are concerned with whether we should adopt an operationalist view of this algorithm, or if the structural core should be furnished with a realistic interpretation. And of course, providing such a realistic interpretation requires that one provide a resolution to the quantum measurement problem.

Giving a realistic interpretation of quantum field theory similarly requires a solution to the quantum measurement problem, but the measurement problem is often conspicuously absent in foundational discussions of the theory. One reason for this is that relativistic constraints raise difficulties for generalizing some solutions to the measurement problem from quantum mechanics to quantum field theory. Another reason is that quantum field theory is often characterized as a theory of scattering.[2] This can be seen from the fact that the basic phenomenological object in the theory is often taken to be the S-matrix which encodes transition amplitudes between prepared incoming states and measured outgoing states, both with determinate particle content.

Draft of 3 July 2020

[†]Department of Philosophy, University of Toronto

[1] Helpful critical discussion of what belongs to the structural core, and what does not, can be found in (Wallace 2019).

[2] The historical reasons for this are discussed in (Blum 2017).

So one might worry that before we even get to the issue of the measurement problem, the formalism for the theory is tinged with operationalism. The structure of the theory is designed to capture the scattering experiments used to test the theory from the outset.

Suppose we are interested in pressing on and attempting to give a realist interpretation of the scattering phenomena that quantum field theory *is* able to describe. We can use the scattering form of Born's rule,

$$Pr(\psi_{\text{out}}|\psi_{\text{in}}) := |\langle\psi_{\text{out}}|S|\psi_{\text{in}}\rangle|^2, \tag{1}$$

to determine the probability of a transition from the state $|\psi_{\text{in}}\rangle$ to the state $|\psi_{\text{out}}\rangle$. On first inspection, this seems to involve essentially the same structural core as non-relativistic quantum mechanics, and to provide an algorithm for predicting the outcomes of experiments which we can go about interpreting. However, the quantum field theoretic algorithm is beset with interpretive challenges of its own that arise before we confront the measurement problem. As a result, much of the interpretive work dedicated to quantum field theory has been concerned with the processes that are required to get the algorithm up and running, and not the interpretation of the algorithm itself.

The interpretive difficulties facing the quantum field theoretic algorithm are diverse. For one, $|\psi_{\text{in}}\rangle$ and $|\psi_{\text{out}}\rangle$ are not states in the physical statespace of the interacting quantum fields involved in the scattering. Rather, they are states in the statespace of free fields. Information about the interacting fields must be gleaned from the perturbative evaluation of the S-matrix element for a particular $|\psi_{\text{in}}\rangle$ and $|\psi_{\text{out}}\rangle$. To do this we sum all of the Feynman diagrams with the appropriate particle content and incoming and outgoing momenta. This perturbative evaluation gives rise to additional obstacles to interpretation. The most widely discussed of these are the ultraviolet divergences that arise from the short-distance and large-momentum regime of the theory. The integrals corresponding to individual diagrams contributing to the probabilities in Eq. (1) are infinite. These ultraviolet divergences necessitate the renormalization of the theory in order to render predictions for the outcomes of experiments finite.[3] Some presentations of the theory give the impression that a properly implemented renormalization procedure is sufficient to get an algorithm up and running that gives probabilities that match the experimental results.

[3]With the development of the renormalization group, the physical need for this process is now well-understood. Quantum field theories are understood as effective theories with an explicitly specified domain of applicability. Recent philosophical literature has begun to address how this approach to understanding the ultraviolet divergences might affect the prospects for realist interpretations of the the theory. For my purposes, the important conclusion that can be drawn from these discussions is that the ultraviolet divergences do not provide an obstacle to realist interpretations of field theory.

In fact, an additional step is required. There is an independent source of infinities that need to be addressed before the algorithm yields finite probabilities. These infrared divergences come from the long-distance and small-momentum regime of the theory, and have received comparatively little attention in the literature. The infrared divergences result from the emission of very low momentum massless particles, and are typically regarded as less conceptually problematic than ultraviolet divergences because there is a reasonably straightforward cancellation mechanism that renders physical observables such as decay rates and cross-sections infrared finite. More precisely, the infrared divergences cancel when we restrict to *measurable* physical quantities. My aim in this paper is to scrutinize the restriction to measurable physical observables that is required to make the cancellation mechanism applicable. It is prima facie plausible that there are physical quantities that are not measurable, but about which there are still facts. For this reason, a restriction to what is measurable is potentially problematic. If one adopts an operationalist interpretation which only countenances those quantities which are measurable as meaningful, such a restriction is unproblematic. However, if one ultimately aspires to provide a realist interpretation, one needs the quantum field theoretic algorithm to be well-defined for all of the physically meaningful quantities, which may not just be the measurable ones. So to ensure that the restriction in question does not amount to a thumb on the operationalist's side of the scale, we need to make sure that we are not restricting beyond the physical matters of fact.

In order to determine whether or not the restriction to measurable physical quantities is an acceptable one, we must analyze the origin of the infrared divergences and the infrared cancellation mechanism in detail. I turn to that task in Section Two. In Section Three I discuss the restriction to measurable physical quantities and I argue that it need not mark a problematic retreat to operationalism. In the fourth section I argue that the infrared divergences from massless particles are a conceptually distinct infrared problem from the one raised by Haag's theorem. The infrared divergences discussed here are more directly relevant for the prospects of providing a realist interpretation of the theory because they bear on the nature of the physical statespace of the theory. Section Five concludes by emphasizing that the infrared regime of quantum field theory contains foundationally significant issues which are important for the project of interpreting the theory.

2. Infrared Cancellation. Early in the development of quantum electrodynamics it was recognized that the infrared problems of classical electrodynamics carried over to quantum field theory. In this latter context, the problems stem from the presence of massless particles. If a massless particle is "soft" in the sense that it has very low momentum, then the emission of such a particle requires very little energy. In the case of quantum electrody-

namics, for example, in processes with outgoing electrons in the final state, the electron is never actually free as we are accustomed to thinking of it. In reality, outgoing electrons emit many soft photons which lead to infrared divergences in the S-matrix element for the process.[4] Closely analogous problems arise in quantum chromodynamics due to the massless gluons, and in quantum theories of gravity involving massless gravitons.

An approach to addressing the infrared divergences was discovered by Bloch and Nordsieck even before the development of covariant perturbation theory for quantum electrodynamics (Bloch and Nordsieck 1937). What they realized was that the infrared divergences from the emission of soft photons are perfectly cancelled by infrared divergences from virtual soft photons. This cancellation mechanism was elaborated in full detail for quantum electrodynamics by Yennie, Frautschi and Suura who showed conclusively that QED can be rendered infrared finite to all orders of perturbation theory (Yennie, Frautschi, and Suura 1961). Weinberg produced a significant simplification of the argument, which also applies to theories with massless gravitons, shortly after (Weinberg 1965). Similar arguments, though more limited in their generality, have also been provided for quantum chromodynamics.[5] The central observation required to induce the cancellation in each case is that any realistic particle detector has some minimum energy threshold. Particles with energy below this threshold will pass through the detector undetected. When S-matrix elements, transition rates, and cross-sections are expressed in a way that accounts for the presence of such a threshold, the infrared divergences can be shown to cancel to all orders.

Suppose we are interested in a QED process with initial state α and final state β containing a total of n incoming and outgoing electrons.[6] The S-matrix element for this process $S_{\beta\alpha}$ requires corrections from the emission of soft photons. Consider the simplest case where a single soft photon is emitted from one of the outgoing electron lines as shown in Fig. 1(a). This yields a correction given by the product of an electron-photon vertex, and an electron propagator with momentum $p+q$, in the limit where $q \to 0$:[7]

$$\left[i(2\pi)^4 e(2p^\mu + q^\mu)\right] \cdot \left[\frac{-i}{(2\pi)^4}\frac{1}{(p+q)^2 + m^2 - i\epsilon}\right] \xrightarrow{q\to 0} \frac{ep^\mu}{p\cdot q - i\epsilon}. \qquad (2)$$

[4] Additional infrared divergences can occur when massless particles move collinearly with the particle from which they were emitted. This class of divergences can be addressed with methods similar to those discussed in this section, though they will not be my focus in this paper.

[5] One important example is provided by the KLN theorem (Kinoshita 1962; Lee and Nauenberg 1964). For helpful discussion see (Muta 1987, Ch. 6).

[6] The argument I present here is a simplified version of the one initially given in (Weinberg 1965) and further elaborated in (Weinberg 1995, Ch. 13).

[7] In taking the limit I have used the freedom to rescale ϵ without changing the sign of the term.

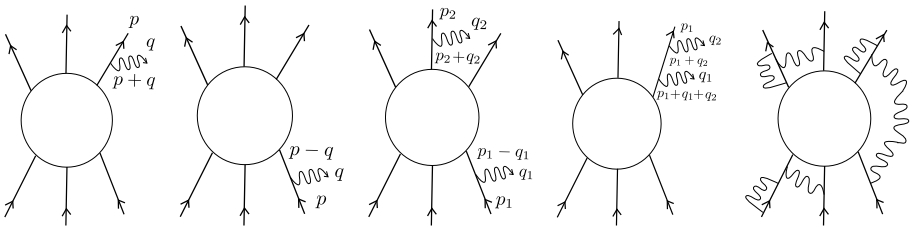


Figure 1: Emission of real soft photons, and exchange of virtual soft photons.

If the photon is emitted from an incoming line rather than an outgoing line, as shown in Fig. 1(b), then the momentum in the additional propagator is $p - q$ and the correction is given by:

$$\left[i(2\pi)^4 e(2p^\mu - q^\mu)\right] \cdot \left[\frac{-i}{(2\pi)^4} \frac{1}{(p-q)^2 + m^2 - i\epsilon}\right] \xrightarrow{q \to 0} \frac{ep^\mu}{-p \cdot q - i\epsilon}. \tag{3}$$

To obtain the correction for the emission of a single soft photon from any of the incoming or outgoing electron lines we must sum over each way the process can happen. If we adopt the convention that $\eta_n = +1$ if the emission is from an outgoing line and $\eta_n = -1$ if it is from an incoming line, this sum can be written compactly as:

$$\sum_n \frac{\eta_n e p_n^\mu}{p_n \cdot q - i\eta_n \epsilon}. \tag{4}$$

If two soft photons are emitted, the correction is given by a product of factors like those we found in Eq. (2) and Eq. (3). For example, if one is emitted from an incoming line and one is emitted from an outgoing line, as in Fig. 1(c), the correction is given by:

$$\left[\frac{ep_2^\mu}{p_2 \cdot q_2 - i\epsilon}\right] \cdot \left[\frac{ep_1^\mu}{-p_1 \cdot q_1 - i\epsilon}\right]. \tag{5}$$

If both electrons are emitted from the same outgoing line, as in Fig. 1(d), then the correction is:

$$\left[\frac{ep_1^\mu}{p_1 \cdot q_2 - i\epsilon}\right] \cdot \left[\frac{ep_1^\mu}{p_1 \cdot (q_1 + q_2) - i\epsilon}\right]. \tag{6}$$

A simple induction[8] shows that the correction for the emission of N soft

[8]See (Weinberg 1995, pp. 538-539).

photons is given by:

$$\prod_{i=1}^{N}\left(\sum_{n}\frac{e\eta_n p_n^{\mu_i}}{p_n\cdot q_i - i\eta_n\epsilon}\right). \tag{7}$$

From this basic relation we can determine the effects of both virtual and real soft photons on $S_{\beta\alpha}$.

To determine the correction from the contribution of soft virtual photons depicted in Fig. 1(e), we must introduce a scale Λ which determines which virtual photons we want to count as soft. Different choices of Λ simply correspond to different choices of what count as radiative corrections, and what count as part of the uncorrected matrix element. We will also be manipulating infrared divergent expressions and so we will introduce an infrared cutoff λ. This cutoff will eventually be removed by taking the $\lambda \to 0$ limit at the end of the calculation.

The correction from a single soft virtual photon can be determined by taking the product of two emitted photon corrections, multiplied by a photon propagator $(-ig_{\mu\nu})/[(2\pi)^4\cdot(q^2-i\epsilon)]$, summing over the polarization indices, and integrating over the soft photon momentum:

$$\int_{\lambda}^{\Lambda} d^4q\, A(q), \tag{8}$$

where,

$$A(q) = \frac{-i}{(2\pi)^4(q^2-i\epsilon)}\cdot\sum_{n,m}\frac{e^2\eta_n\eta_m(p_n\cdot p_m)}{(p_m n\cdot q - i\eta_n\epsilon)(-p_m\cdot q - i\eta_m\epsilon)]}. \tag{9}$$

To obtain the correction from N virtual soft photons, we take the product of N such factors, and divide by factors of $N!$ to account for possible permutations of where the lines attach, and (2^N) to account for interchanges of the two ends of the line. This gives,

$$\frac{1}{N!}\left[\frac{1}{2}\int_{\lambda}^{\Lambda} d^4q\, A(q)\right]^N, \tag{10}$$

and thus, when we sum over N and use the fact that $\exp(x) = \sum_N x^N/N!$ we find that,

$$S^{\lambda}_{\beta\alpha} = S^{\Lambda}_{\beta\alpha}\exp\left(\frac{1}{2}\int_{\lambda}^{\Lambda} d^4q\, A(q)\right). \tag{11}$$

$S^{\Lambda}_{\beta\alpha}$ is the S-matrix element with no virtual photon exchange with momentum less than Λ included. $S^{\lambda}_{\beta\alpha}$ is the S-matrix element corrected to include virtual soft photon exchange with momentum greater than λ but less than Λ. The

rate for the process is then given by the matrix element squared:

$$\Gamma^{\lambda}_{\beta\alpha} = |S^{\lambda}_{\beta\alpha}|^2 = |S^{\Lambda}_{\beta\alpha}|^2 \exp\left(\int_{\lambda}^{\Lambda} d^4q A(q)\right) = \Gamma^{\Lambda}_{\beta\alpha} \exp\left(\int_{\lambda}^{\Lambda} d^4q A(q)\right). \quad (12)$$

Weinberg shows that the integral in the exponential yields:

$$\int_{\lambda}^{\Lambda} d^4q A(q) = -A \ln\left(\frac{\Lambda}{\lambda}\right), \quad (13)$$

where,

$$A = \frac{-1}{8\pi} \sum_{n,m} \frac{e^2 \eta_n \eta_m}{\beta_{nm}} \ln\left(\frac{1+\beta_{nm}}{1-\beta_{nm}}\right) \quad \text{and} \quad \beta_{nm} = \left[1 - \frac{m_e^4}{(p_n \cdot p_m)^2}\right]^{1/2}. \quad (14)$$

Inserting Eq. (13) into Eq. (12), and using familiar properties of exponentials and logarithms, we find that:

$$\Gamma^{\lambda}_{\beta\alpha} = \Gamma^{\Lambda}_{\beta\alpha} \exp\left(-A \ln\left(\frac{\Lambda}{\lambda}\right)\right) = \Gamma^{\Lambda}_{\beta\alpha} \left[\exp\left(\ln\left(\frac{\lambda}{\Lambda}\right)\right)\right]^{A} = \Gamma^{\Lambda}_{\beta\alpha} \left(\frac{\lambda}{\Lambda}\right)^{A}. \quad (15)$$

This provides a complete statement of the correction to the rate from virtual soft photons. In the limit where $\lambda \to 0$ we see that the rate $\Gamma^{\lambda}_{\beta\alpha}$ vanishes. This is the result of exponentiating $\ln(\Lambda/\lambda)$ which is divergent in the $\lambda \to 0$ limit.

The virtual soft photon divergences leading to this unphysical vanishing of the rate are cancelled by divergences from real photon emission. More precisely, this cancellation can be seen to apply to all orders of perturbation theory when the total rate, including all radiative corrections, is expressed in terms of the resolution of the detector used to measure the real soft photons. Weinberg explains the restriction as follows:

> The resolution of the infrared divergence problem ... is found in the observation that *it is not really possible to measure the rate* $\Gamma_{\beta\alpha}$ for a reaction $\alpha \to \beta$ involving definite numbers of photons and charged particles, because photons of very low energy can always escape undetected. *What can be measured* is the rate $\Gamma_{\beta\alpha}(E, E_T)$ for such a reaction to take place with no unobserved photon having an energy greater than some small quantity E, and with not more than some small total energy E_T going into any number of unobserved photons. (Weinberg 1995, pp. 544-545, my emphasis)

This restriction to the measurable quantity $\Gamma_{\beta\alpha}(E, E_T)$ in order to render the rate infrared finite requires careful analysis. I will turn to that task in Section

Three. The remainder of this section completes the demonstration that if one makes this restriction, then the infrared divergences cancel.

In order to calculate the correction from the emission of N real soft photons, with momenta $q_1, \ldots, q_N$, each term in Eq. (7) must be multiplied by the appropriate coefficient function,[9]

$$\frac{\epsilon^*_\mu(\mathbf{q_i}, h_i)}{(2\pi)^{3/2}(2|\mathbf{q_i}|)^{1/2}}. \tag{16}$$

This yields the following expression for the matrix element $S^\lambda_{\beta\alpha}(q_1, q_2, \ldots, q_N)$, which includes the contributions of both the virtual soft photons and the N real emitted soft photons:

$$S^\lambda_{\beta\alpha}(q_1, q_2, \ldots, q_N) = S^\lambda_{\beta\alpha} \prod_{i=1}^{N} \frac{1}{(2\pi)^{3/2}(2|\mathbf{q}_i|)^{1/2}} \cdot \sum_n \frac{\eta_n e(p_n \cdot \epsilon^*(\mathbf{q}_i, h_i))}{(p_n \cdot q_i)}, \tag{17}$$

where $S^\lambda_{\beta\alpha}$ is as given in Eq. (11). The differential rate for the emission of N soft photons into the volume of momentum space $\prod_i d^3\, q_i$, is given by squaring Eq. (17), summing over the helicities, and multiplying by $\prod_i d^3\, q_i$ which gives:

$$d\,\Gamma^\lambda_{\beta\alpha}(q_1, q_2, \ldots, q_N) = \Gamma^\lambda_{\beta\alpha} \prod_{i=1}^{N} \frac{d^3\, q_i}{(2\pi)^3(2|\mathbf{q}_i|)} \cdot \sum_{nm} \frac{\eta_n \eta_m e^2 (p_n \cdot p_m)}{(p_n \cdot q_i)(p_m \cdot q_i)} \tag{18}$$

Integrating over the direction of photon propagation yields the differential rate for the emission of N soft photons with energies $\omega_1, \ldots, \omega_N$:

$$d\,\Gamma^\lambda_{\beta\alpha}(\omega_1, \omega_2, \ldots, \omega_N) = \Gamma^\lambda_{\beta\alpha} A^N \frac{d\omega_1}{\omega_1} \frac{d\omega_2}{\omega_2} \cdots \frac{d\omega_N}{\omega_N} \tag{19}$$

where the factor A is as defined in Eq. (14). Note that if we were to integrate Eq. (19) over the emitted energies of the photons, we would produce logarithmic divergences from the $\omega \to 0$ end of the integrations. However, the imposition of the infrared cutoff λ ensures that the expressions are regulated. If we were to remove the regulator at this stage of the calculation, the cancellation mechanism would not do its job, and we would not arrive at a sensible physical rate at the end of the calculation.

In order to arrive at a final expression for the rate, the integration over photon energies must be done respecting the constraints described in the quotation of Weinberg above. In particular, the unobserved photons must each have energy below the detector threshold and above the infrared cutoff, $E \geq \omega_i \geq \lambda$, and the total energy of all of the unobserved photons must not

[9] In this expression, ϵ is a polarization vector and h is the helicity.

be greater than E_T, $\sum_i \omega_i \leq E_T$:

$$\Gamma^{\lambda}_{\beta\alpha}(E, E_T) = \Gamma^{\lambda}_{\beta\alpha} \sum_{N=0}^{\infty} \frac{A^N}{N!} \int_{E \geq \omega_i \geq \lambda,\, \sum_i \omega_i \leq E_T} \prod_{i=1}^{N} \frac{d\,\omega_i}{\omega_i}, \tag{20}$$

The integration subject to these restrictions gives:[10]

$$\Gamma^{\lambda}_{\beta\alpha}(E, E_T) = \left(\frac{E}{\lambda}\right)^A \Gamma^{\lambda}_{\beta\alpha}. \tag{21}$$

The cancellation of the infrared divergences is achieved by inserting the expression in Eq. (15) for $\Gamma^{\lambda}_{\beta\alpha}$ into Eq. (21). This combines all corrections from real and virtual photons into an expression for $\Gamma^{\lambda}_{\beta\alpha}(E, E_T)$:

$$\Gamma^{\lambda}_{\beta\alpha}(E, E_T) = \left(\frac{E}{\lambda}\right)^A \Gamma^{\lambda}_{\beta\alpha} = \left(\frac{E}{\lambda}\right)^A \left(\frac{\lambda}{\Lambda}\right)^A \Gamma^{\Lambda}_{\beta\alpha} = \left(\frac{E}{\Lambda}\right)^A \Gamma^{\Lambda}_{\beta\alpha}. \tag{22}$$

Note that the factors of λ cancelled each other, and so we can take $\lambda \to 0$ to obtain:

$$\Gamma_{\beta\alpha}(E, E_T) = \left(\frac{E}{\Lambda}\right)^A \Gamma^{\Lambda}_{\beta\alpha}. \tag{23}$$

Thus, when we account for both soft virtual photon exchange and real soft photon emission, the rate becomes independent of λ and is infrared finite. The procedure used to achieve this result does, however, introduce a dependence on the detector resolution, E.

The subsequent literature adopts a distinction between exclusive and inclusive quantities.[11] Exclusive quantities stipulate the exact contents of the incoming and outgoing states. For example, in an exclusive cross-section one might demand that there are exactly three electrons and no other particles, even if the other particles are not detected. Inclusive quantities stipulate part of the contents of the final state, but they also account for the possibility that there are other particles in the final state. The rate in Eq. (23) provides an example of an inclusive quantity. We have stipulated that there are a total of n incoming and outgoing electron lines, but we have also accounted for the emission of an arbitrary number of undetected soft photons each with energy less than E and with total energy less than E_T. At particle accelerators, attention is often restricted to such inclusive quantities, and the it is the justification for this to which we now turn our attention.

[10] I have omitted an overall factor resulting from the integration which is close to 1 in the circumstances we are interested in analyzing.

[11] As far as I have been able to determine, this distinction originates from (Feynman 1969).

3. Measurement. The apparent need to restrict to measurable physical quantities has arisen in other contexts during the development of quantum field theory. Early in the development of the theory, Bohr and Rosenfeld argued that the value of the field at a point was not a measurable quantity, but that the average value of the field over a small spacetime region was measurable (Bohr and Rosenfeld 1933; Bohr and Rosenfeld 1950). It was later realized that field operators could not be mathematically defined at points of spacetime, and that instead they had to be represented as operator-valued distributions which are well-defined only as integrations against test functions of compact support on small regions of spacetime.[12] When it was realized that the mathematical definition of the theory became ill-defined for associations of operators with points, a number of figures suggested that this should be intepreted as resulting from the fact that such quantities were unmeasurable.[13] If one adopts the additional assumption that unmeasurable quantities are not meaningful, then the ill-definedness of field operators at points becomes unproblematic: there is no physically meaningful quantity for the ill-defined field operators to correspond to.

Similar reasoning has been employed to address other ill-defined quantities from the ultraviolet regime. Empirically interesting field theories are ultraviolet divergent and require renormalization. This process involves recognizing that some parameters in the lagrangian such as the bare mass and the bare charge are infinite and introducing counterterms to cancel the infinities and re-express the theory in terms of measurable parameters such as the dressed mass and charge. In response to this situation one frequently encounters the claim that bare parameters in the Lagrangian are unmeasurable. To take just one example, Srednicki explains that "It may be disturbing to have a parameter in the Lagrangian that is formally infinite. However, such parameters are not directly measurable, and so need not obey our preconceptions about their magnitudes" (Srednicki 2007, p. 67).[14] Once again, we encounter the view that only those quantities that are measurable are required to be meaningful.

Compare this to the reasoning Weinberg offered in the previous section. The rate $\Gamma^{\lambda}_{\beta\alpha}$ is infrared divergent in the limit where $\lambda \to 0$, but it is unmeasurable. The measurable rate $\Gamma^{\lambda}_{\beta\alpha}(E, E_T)$ is infrared finite to all orders of perturbation theory in the $\lambda \to 0$ limit. The justification for the need to make this restriction in order to arrive at infrared finite quantities, when one is explicitly articulated, is that any real physical detector has some finite energy resolution and particles with energy below that threshold will not be

[12] This came to be understood in stages, with the conclusive theorem provided in (Wightman 1964).

[13] (Friedrichs 1951; Cook 1953)

[14] Similar claims can be found in (Peskin and Schroeder 1995, p. 315) and (Itzykson and Zuber 2012, p. 319), and in many other accounts of the rationale underlying renormalization.

registered in the detector. Thus, Weinberg's demonstration establishes that quantum field theory provides well-defined values for all of those observables that are physically measurable and most discussions of this issue leave off here.[15]

Absent additional argumentation, I think that this amounts to a problematic retreat to operationalism. To be clear, my concern is not with operationalism as an account of meaning in general. I am open to the possibility that operationalism provides a compelling account of meaning in at least some cases. What is problematic in this case is that the justification for the restriction to measurable quantities relies on the stronger claim that *only* those quantities that are measurable are physically meaningful. Suppose this stronger claim were true. Then the demonstration that the field theoretic expressions for the measurable observables are well-defined amounts to a demonstration that the field theoretic expressions for every physically meaningful quantity is well-defined. If the stronger claim is not true, and there are physically meaningful quantities that are not measurable, then the demonstration that the measurable quantities are well-defined does not go far enough to establish that the theory adequately accounts for all of the meaningful quantities.

To determine whether or not the restriction to measurable quantities in the infrared case is problematic, we need to know whether or not failures of measurability stand in direct correspondence with failures of meaningfulness. For this reason, each proposed restriction to measurable quantities requires its own analysis, as each involves distinct physical limitations on what is measurable. While I believe that both of the ultraviolet cases introduced above merit further attention of their own, here I will restrict attention to the infrared case as that is my central concern in this paper.

Suppose we simply grant that every physical detector will have some threshold E such that particles with energy less than E will not be detected.[16] Note that quantities like cross-sections and rates are defined with respect to a particular collection of incoming particles, and a particular collection of detected outgoing particles. However, for a given incoming state, α, the dynamics of the theory will yield an outgoing state which is a superposition with indeterminate particle content, including an indeterminate number of electrons, hard photons, and soft photons with energy below the detection threshold. It is only upon measurement that the outgoing state becomes

[15] Essentially the same justification can be found throughout the physics literature. See, for example, (Brown 1992, pp. 490-491), (Duncan 2012, p. 719, p. 723, p. 728), (Itzykson and Zuber 2012, p. 173, p. 354), (Peskin and Schroeder 1995, pp. 200-202), (Schweber 2011, p. 549), and (Srednicki 2007, pp. 157-158).

[16] This claim is often asserted without argument. Establishing its validity would require a detailed analysis of the physical nature of the detector and its coupling to the measured particle. I am grateful to Jeff Barrett for discussion of this point.

one with the determinate particle content as we assumed β to have. And, of course, how one conceives of this process of becoming a state with determinate particle content depends on how one prefers to resolve the quantum measurement problem.

In computing the rate $\Gamma_{\beta\alpha}(E, E_T)$ we assumed that this measurement process yields a specific number of electrons and no hard photons in the final state. If there were hard photons, or a different number of electrons, we would need to compute the rate for a different process. Given that there are outgoing electrons in the final state, there are also soft photons which were not detected. So the justification relied on here is not that there is no photon detector that can detect arbitrarily soft photons and hence quantities involving them are meaningless. Rather, every measurement that is done has some energy resolution, and we need to account for the fact that given the particular measurement that has been executed, there can be soft photons below that resolution.

This shows why it necessary to express the physical quantities in terms of the detector resolution, E. For a given incoming state, there are distinct possible outgoing states. By selecting a specific β, we have not done quite enough to specify which part of the statespace the measurement is a projection onto. By specifying E, we condition on which kinds of soft radiation can be undetected in the final state. For a different detector energy resolution E', different kinds of unobserved soft radiation states are possible, as are different alternatives to β. The need to restrict to what is measurable is not a retreat to operationalism. Rather, the presence of the energy resolution is an articulation of the precise nature of the question we are asking about the outgoing state by executing the particular measuring process that we chose to execute.

4. The Connection to Haag's Theorem. In their appraisal of the philosophical significance of Haag's theorem, Earman and Fraser make several references to infrared divergences (Earman and Fraser 2006). They claim, for example, that "In the physicists' lingo, the move from one inequivalent representation to another is marked by divergences. Haag's theorem is concerned with infrared divergences that are associated with Euclidean invariance and the infinite volume of space (Earman and Fraser 2006, p. 319)". They also note the infrared divergences can be tamed by imposing some form of infrared regulator.[17] The imposition of an infrared regulator can cure more than one kind of infrared pathology, and caution is required here in order not to run together two conceptually distinct issues.

The interaction picture is a formal intermediary between the the Schrödinger picture and the Heisenberg picture which is often employed as a calculational

[17]The regulators they consider are the compactification of space, and the restriction of the theory to bounded regions of spacetime (Earman and Fraser 2006, p. 319, 323, 330).

tool to facilitate the perturbative evaluation of observables. It postulates the existence of a global unitary transformation connecting the free and interacting Hilbert spaces. Haag's theorem shows that this transformation does not exist and that these spaces are unitarily inequivalent. Thus, the interaction picture is predicated on an inconsistent set of assumptions. Miller has provided an account of how perturbative calculations that employ the interaction picture can be empirically successful despite this apparent inconsistency (Miller 2016). The imposition of an infrared regulator renders some of the assumptions of the theorem false. This undercuts the threat to the empirical success of the theory from Haag's theorem, but it leaves questions about the well-definedness of the interaction picture in the limit where the regulator is removed.

Infrared divergences from soft massless particles raise a more serious worry about the infrared regime of quantum field theory than the one implicated in Haag's theorem. The infrared cancellation results are sufficient to assuage worries about how it can be that theories with infrared divergences are still empirically successful. However, because of the presence of the soft massless particles, free electron states with distinct momenta are unitarily inequivalent to one another.[18] As such, this class of infrared divergences call into question the well-definedness of the physical state spaces of theories like quantum electrodynamics. For this reason, I think they are rightly regarded as a symptom of more serious conceptual problem than Haag's theorem, which only undermines a method for extracting predictions from the theory. The challenge from the soft massless particles is a serious one for interpreters of quantum field theory and it is one which in my view requires significant further attention.[19]

5. Conclusion. I have argued that the need to express physical quantities in terms of the energy resolution of a detector does not mark a problematic retreat to operationalism. As in the case of the ultraviolet divergences, the infrared divergences can be understood physically. With a properly implemented renormalization scheme and infrared cancellation mechanism in place, the algorithm of quantum field theory provides finite expressions for physical observables. Thus, the infrared divergences, like the ultraviolet divergences, are not ultimately an obstacle to realist interpretations of the theory. The infrared regime of the theory is fraught with conceptual issues which bear directly on the issue of how one might go about producing such an interpretation, and very much warrants further attention from a foundational perspective.

[18] For discussion see (Duncan 2012, pp. 722-723) or (Buchholz 1982).

[19] Perhaps the first philosopher to approach this problem is Ruetsche, who has suggested that coherent state representations may play an important role in understanding these issues (Ruetsche 2012, pp. 245-246).

References

Bloch, F. and A. Nordsieck (1937). Note on the Radiation Field of the Electron. *Phys. Rev. 52*, 54–59.

Blum, A. S. (2017). The state is not abolished, it withers away: How quantum field theory became a theory of scattering. *Stud. Hist. Phil. Sci. B60*, 46–80.

Bohr, N. and L. Rosenfeld (1933). Zur frage der messbarkeit der elektrimagnetischen feldgrössen. *Det Kgl. Danske Videnskabernes Selskab. Mathematisk-fysiske Meddelelser 12*, 1–65.

Bohr, N. and L. Rosenfeld (1950). Field and charge measurements in quantum electrodynamics. *Phys. Rev. 78*, 794–798.

Brown, L. S. (1992). *Quantum Field Theory*. Cambridge University Press.

Buchholz, D. (1982). The physical state space of quantum electrodynamics. *Comm. Math. Phys. 85*(1), 49–71.

Cook, J. M. (1953). The mathematics of second quantization. *Transactions of the American Mathematical Society 74*(2), 222–245.

Duncan, A. (2012). *The conceptual framework of quantum field theory*. Oxford: Oxford Univ. Press.

Earman, J. and D. Fraser (2006). Haag's Theorem and its Implications for the Foundations of Quantum Field Theory. *Erkenntnis 64*.

Feynman, R. P. (1969). Very high-energy collisions of hadrons. *Phys. Rev. Lett. 23*, 1415–1417.

Friedrichs, K. O. (1951). Mathematical aspects of the quantum theory of fields parts i and ii. *Communications on Pure and Applied Mathematics 4*(2-3), 161–224.

Itzykson, C. and J. Zuber (2012). *Quantum Field Theory*. Dover Books on Physics. Dover Publications.

Kinoshita, T. (1962). Mass singularities of Feynman amplitudes. *J. Math. Phys. 3*, 650–677.

Lee, T. D. and M. Nauenberg (1964). Degenerate Systems and Mass Singularities. *Phys. Rev. 133*, B1549–B1562.

Miller, M. E. (2016). Haag's Theorem, Apparent Inconsistency, and the Empirical Adequacy of Quantum Field Theory. *The British Journal for the Philosophy of Science 69*(3), 801–820.

Muta, T. (1987). Foundations of quantum chromodynamics: An Introduction to perturbative methods in gauge theories. *World Sci. Lect. Notes Phys. 5*, 1–409.

Peskin, M. and D. Schroeder (1995). *An Introduction To Quantum Field Theory*. Frontiers in Physics. Avalon Publishing.

Ruetsche, L. (2012). *Interpreting Quantum Theories*. Oxford: Oxford Univ. Press.

Schweber, S. (2011). *An Introduction to Relativistic Quantum Field Theory*. Dover Publications.

Srednicki, M. (2007). *Quantum Field Theory*. Cambridge University Press.

Wallace, D. (2019). What is orthodox quantum mechanics? In A. Cordero (Ed.), *Philosophers Look at Quantum Mechanics*. Springer Verlag.

Weinberg, S. (1965). Infrared photons and gravitons. *Phys. Rev. 140*, B516–B524.

Weinberg, S. (1995). *The Quantum Theory of Fields*, Volume 1. Cambridge University Press.

Wightman, A. S. (1964). La théorie quantique locale et la théorie quantique des champs. In *Annales de l'IHP Physique théorique*, Volume 1, pp. 403–420.

Yennie, D. R., S. C. Frautschi, and H. Suura (1961). The infrared divergence phenomena and high-energy processes. *Annals Phys. 13*, 379–452.

Engineering roles and identities in the scientific community: toward participatory justice[1]

V. Pronskikh[2]
Fermi National Accelerator Laboratory, Batavia, IL 60510-5011, USA[3]

Abstract

This paper seeks to examine the roles and identities of engineers constituting one of the fundamental, but a completely indescribable community in modern big science with particle accelerators. Large communities of accelerator and detector specialists, which replaced experimenters and instrumentalists of the middle of the last century, themselves exhibit a complex structure and are divided. However, this division is in turn grounded on the division of those whose activities focus on the phenomena of nature considered independent of human beings and those who design processes and phenomena of an artificial, technical nature. Nevertheless, in terms of their modus operandi and identity, the kinship between engineers and experimental scientists is considerable. I argue that such exclusion of the engineering community from epistemic practices can serve as an example of participatory injustice. As one of the ways to transcend participatory injustice, I suggest that the communities should be encouraged to work together in epistemically tantamount roles while structural hindrances to the mobility between communities need to be alleviated.

Keywords: engineering, high-energy physics, communities, roles, mobility, participatory justice

Introduction

Contemporary fundamental science has a number of significant contrasts with the science of the early 20th century. Having become a complex social institution in its very structure, it demanded the eliciting and scrutinizing of the communities that make up research teams and the features of their interaction. Several outstanding studies have been undertaken by a number of historians, philosophers, and sociologists of science (Galison 1987; Pickering 1988; Collins 2002, Hoddeson et al. 2008, Knorr-Cetina 1999, Traweek 1988, Latour and Woolgar 1979). The objective of this work is to examine the phenomenon of the big science community in its development and the influence of the rise of an elementary particle accelerator and a complex elementary particle detector in it on the amplification of the social structure, the deepening of the epistemic division

[1] Talk accepted for presentation at PSA2020: The 27th Biennial Meeting of the Philosophy of Science Association (Baltimore, MD; 18-22 Nov 2020).
[2] vpronskikh@gmail.com
[3] Fermi National Accelerator Laboratory is operated by the Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

of labor and the need for lengthy engineering activity at the stage of preparation of experiments. To accomplish that task, it becomes necessary to clarify the roles of members of scientific communities and the dynamics of changes in their structure, to discuss the difficulties of classification and identities of community members, as well as issues that arise at the present stage in the relationship of scientific and engineering activities and possible ways to resolve them.

Notably, the rise of the accelerator in the 1930s–1940s was the first milestone in the development of a modern complex physics experiment and the complication of the structure of communities associated with large-scale experiments in high-energy physics. First, the emergence of such a large and complex device as an accelerator led to the appearance of accelerator physicists (and engineers) as scientific and technical specializations. Their task comprised the design of the accelerator, the calculation and optimization of its parameters, as well as ensuring its operation (providing particle beams accelerated to the required energies and intensities, to the community of experimentalists). The creation of the accelerator not only led to the spatial separation of the theoretician working silently in their Ivory Tower from the experimentalist, who now had to spend most of their time in the experimental halls near the accelerator, where their installation was established. In addition to the communities of experimentalists and theorists, with the beginning of experiments on accelerators, a community of accelerator specialists emerged, engaged in the creation and maintenance of the accelerator machine. Galison (1987) also introduced a community of instrumentalists involved in the creation of scientific instruments and installations, and formally accelerator scientists could be classified as such a community because the accelerator is a technically sophisticated device whose operating principles are essentially based on classical electrodynamics. However, in such a case, it would turn out to be very heterogeneous because the expertise of the accelerator researcher and the instrumentalist, who builds, for example, a particle detector, will differ.

Beginning from the 1970s and finally by the beginning of the 1990s, a detector—a device in which particles born in collisions of a beam are detected, and their characteristics are identified by measurement, ended up to be so tangled and universalized that it morphed into the central object of the experiment that in many respects began to designate the long-term directions of research in accelerator laboratories (Hoddeson 2008). The structure of the corresponding communities began to change accordingly. Now, experimenters became engaged in detector calculations and design for a long time, taking on some of the tasks that were previously assigned to instrumentalists (in particular, engineers), then undertaking measurements with it and analyzing the data for an equally long time. After a series of measurements, they often continued to improve the design of the detectors and their components, returning to the engineering kind of work, again measuring and analyzing until the range of tasks that can be solved with this type of detector and accelerator capabilities was exhausted. All together, it took dozens of years, sometimes the whole conscious life of the experimentalist of this particular type of detector or the same detector. However, the communities of accelerator and detector researchers themselves are also heterogeneous. We shall consider their structure in more detail.

Accelerator and detector researchers

Starting from the 1990s, one can assume that instead of communities of theorists, experimenters, and instrumentalists, in high-energy physics, one should talk about communities of theorists,

accelerator, and detector researchers. The structure of the theoretician community has not changed much, while other communities have undergone the greatest changes since the first third of the 20th century. The community of accelerator specialists now builds and maintains the accelerator, and the community of detector specialists—the detector. Instrumentalists can now be considered a subset of detector and accelerator specialists, and the dividing line between them and detector experimental experts is rather blurred. At the stage of creating the setup, the distinction between experimental scientists and nonscientists is such that, although they both study processes in the detector on computational models, the former focus more on aspects related to future searches for a useful signal, its reconstruction (reconstruction of events occurring in the detector by triggering numerous sensors), while the second—on aspects related to ensuring the overall operability of the installation. The second difference between them becomes evident after the beginning of measurements at the setup, when the scientists participate in the data acquisition, and then enter on processing and analysis of the data, while the instrumentalists set about the creation of other installations and instruments.

Until about the 30s of the 20th century, the role of an instrumentalist did not exist because their functions were divided between the experimentalist and engineer as follows: the experimentalist formulated the technical requirements for the device (installation) to the engineer (industry) in the form of a set of requirements and those independently manufactured the device, most of which was standard and serial. Then, the experimentalist performed measurements on the setup, performing its adjustment as necessary, as well as data analysis, which was quite simple and not requiring separate education and specialization. With the birth of Big Science and the resulting complexity and uniqueness of the installations, the technical design specifications are becoming a joint product of experimentalists and toolmakers, resulting from a compromise and trade-off between a multitude of installation requirements. In this sense, the instrumentalist (and nowadays, the detector and accelerator researchers) is a transitional type between the engineer and the experimenter, and their own instrumentalists appear in both the accelerator and detector communities.

Roles and specializations in megascience

Each of these communities now become heterogeneous (see Table 1). Accelerator specialists also began to be divided into theorists (calculators) performing computational modeling of the particle acceleration, experimentalists who conducted experimental measurements of the developed accelerator assemblies to help create its technical theories, as well as engineers manufacturing the accelerator assemblies and performing their tuning and adjustment. Detector scientists are divided by the type of detector unit, which they simulate and build, and then support during the measurements and data which they own for analysis after the experiments. With the advent of the era of complex hybrid detectors, the detector began to consist of several complex, but heterogeneous system units, for example, such as a time-of-flight system, calorimeter, tracker, or shielding against the cosmic background. Each of these units, from its design period until the completion of the experiment, was under the responsibility of a certain group of experimentalists, which assumed both a number of technical issues and a physical interpretation of the data harvested from it. These groups together form a community of detector scientists. Engineers are entrusted with the development of installations, accelerators, and their units according to the technical specifications, maintenance of installations, accelerators, and software.

Area	Role	Specialization
Accelerator	Accelerator Physicist	Accelerator Theorist
		Accelerator experimentalist
	Engineer	Accelerator Engineer
Detector	Detector physicist	Unit scientist
		Data measurer
		Data analyst
	Engineer	Unit engineer
Phenomena theory	Theorist	Model developer
		Developer of calculation tools and methods
Computing	Programmer	
		System Tools Developer
	Engineer	Software Maintenance

Table 1. The structure and functions of communities in high-energy physics.

Experimentalists' identity and engineering

To grasp the social processes in the high-energy physics laboratory, the nature of community interactions, their similarities, and dissimilarities, it is necessary to elucidate the identities of their members. As one of the signs of the experimenter's "identity shift," characteristic of the period of the 1970s, the awkwardness felt by the experimental physicist toward others (including the engineer) in the laboratory began to be noted (Galison 1997, 5). This "identity shift," was ascribed to the fact that the very nature of experimentation has changed: if earlier the experimenter's work was unambiguously associated with the design of the installation, the development of experimental procedures, the application of these procedures, the recording of results, and their theoretical analysis, then later the experimenter was considered to be the one who only analyzes the data harvest hiding behind the monitor a long distance away from the installation threshing mill. Hence, it became impossible to have a single view of what can be considered experimentation (which is reflected in Table 1). Another distinctive feature of this stage of the development of science can be considered the complex contradiction between the experimentalist and engineer on the one hand, and the productive tension between the experimentalist and theorist on the other (Galison 1997, 5).

For the sake of our analysis, in the above claims, we highlight the following central narratives: 1) the contradictions between the experimenter and other specialists of the scientific laboratory (theorist and engineer) that have been growing since the 1970s and 2) the emergence of experimentalists who were not engaged in the activities previously considered traditional, such as

creating a facility and experimental procedures for it. These observations appear to be based on the following premises.

First, the community of experimenters implicitly related to detector specialists (as reflected in Table 1) is not uniform but covers a wide range of activities. As was noted, experimentation begins to shift toward data analysis. In practice, it is often believed that, because data analysis constitutes an interpretation of the processes occurring in the detector, in terms of high-level theories, bearing upon the language of instrumental theories in which the principles of the functioning of detectors are rooted, the outcome of this procedure is what directly becomes the experimental result. Because obtaining the measurement result is the experiment's main aim, therefore, the experimentalist, first of all, can be deemed the one whose activity immediately delivers the result, that is, an analysis of detector data.

Notably, most of the participants in the analysis of data (experimentalists) in the period preceding the acquisition and analysis of data are also engaged in the creation of detectors and procedures for them. Thus, the scope of their activity also partially covers that which belongs to the expertise of instrumentalists (who may also be engineers)—the creation of instruments. On the one hand, considering the fact that the creation of the device and the corresponding procedures take a long time (years and even tens of years), the experimenter had to devote a lot of time to the activities that are very close to engineering ones, *de facto* becoming a highly professional engineer. This may raise a legitimate question, why is one of them identified as a detector physicist (experimentalist) and the other as an engineer when their work and professional expertise are so similar? On the other hand, the epistemic distinctions in the nature of their work also blurred: the experimental search, in addition to being carried out, as before, in terms of the dominant theories, was increasingly guided and determined by these theories; the discovery of new phenomena was increasingly dependent on the development of high-level theories. Thus, experimentation became more and more tangibly the construction of theoretical natural objects, which was closer to the work of the designer than before. This became especially pronounced during the period of success of the Standard Model in elementary particle physics, which predicted many particles that were subsequently measured experimentally. This could not but affect the identity of the experimentalist, as well as the perception of the experimenter by the engineer as a theoretician.

The shift of identity, therefore, arose in connection with the need for lengthy engineering work for the experimentalist (because the creation of the installations took a long time and was not serial due to their uniqueness), on the one hand, and the increased constructiveness of the experimentation itself, on the other. This entailed the actual blurring of the lines between the nature of the work of the experimentalist and engineer, which was initially opposed much due to distinct educational trajectories, which are also linked by the mass consciousness with the level of possible scientific horizons and achievable professional competencies.

Why is a scientist more prestigious than an engineer?

Joseph Martin (Martin 2017) has recently argued that prestige of different social and epistemic groups in science, for example, particle physics and applied science (or engineering), is asymmetric (the latter being the least prestigious). The question of what constitutes an engineer's identity requires, first and foremost, an answer to two interrelated questions: Who is an engineer and what features make a person an engineer? Modern literature on identity theory distinguishes cognitive (internal) and social (external) identity (Anderson 2010; Wenger 1998). In the case of an engineer, the former part is predicated upon what they know about the profession, how they

understands their role, what they want (may want) or do not want (may not want) to know professionally. This part is set not only by the engineer's personal cognitive peculiarities but also by the interrelation of their professional role as an engineer with the whole variety of other roles their play in life. The latter part, the social one, is not the traits that an engineer acquires in the course of professional practice, but those shaped by their membership in a social group. The latter part is formed by the community as a result of belonging to it, as well as by other communities and society as a whole, with which the engineer interacts and in which the attitude toward engineers as a social group is defined. This view raises the question of the stability of such an identity and the need for such conditions in the formation of the so-called engineering identity in general. Speaking of engineers as a social group, it becomes necessary to distinguish between an engineer, as the holder of engineering education, and an engineer as a performer of the role of an engineer.

To establish what constitutes an engineer's identity in science, one has to answer the initial question of what features and functions make one an engineer. In the practice of scientific laboratories in basic science (for example, elementary particle physics), the role of an engineer implies working with complex technical systems, but first, we will clarify what the disparities between engineers and scientists boil down to. Historically, these distinctions are rooted in an understanding of the very nature of the activities of these communities and the goals for which these activities, namely scientific and technical research, are oriented. Most approaches to distinguishing science and technology in one way or another reflect the Aristotelian distinction between ἐπιστήμη (episteme) as knowledge, understanding, or cognition and τέχνη (techne) as craft or practical art. The first, according to Aristotle, is a theoretical knowledge of eternal and universal things that exist by virtue of their necessity; the second is the creation of transient and perishable, i.e., practical things. From here originates the ontological distinction between "knowledge of what" and "knowledge of how," knowledge of the true (first) and useful (second). In this regard, a "dichotomy of intellectual status" arose in science and society (Boon 2011, p. 63): higher status of science and lower of technology. At the same time, several authors point out the difficulties of discerning scientific and technical knowledge in modern science, and also advocate the possibility of considering them as either including one another (technical includes epistemic), or even as independent of each other (Boon 2011).

Nonanalyzing data detector and accelerator researchers can be classified as technically oriented scientists (except when they are studying new phenomena during the development of instruments). Engineers (they are not included in the classification (Galison 1987) because they are not classified as scientists) are not engaged in the science of independent development of new devices or study of new phenomena in technical systems, but operate and establish such systems or develop in accordance with the terms of reference, which are formulated by scientists. An engineer who is developing a new system or exploring it according to our view should be classified as a scientist. Thus, the physics of high energies retains the basis of the Aristotelian dichotomy of epistemic and technical which is reflected in the hierarchy of activities and communities in science from pure theorizing about natural phenomena (epistemics) down to technology applications (engineers). Signs of dichotomy remain, however, because even in mixed, intermediate cases, such as those of detector experimental physicists, their activities are clearly divided into two types of roles: the design of the device (for example, the tracker) is technical, its operation is also technical, and the analysis of data with the formulation of theoretical statements is scientific (epistemic). The same applies to the detector scientist, in whose work the epistemic part (the study of new natural phenomena suitable for creating new devices) and the technical (construction, design, and

operation of devices) can clearly be traced. At the poles are theorists, all of whose roles are epistemic and engineers, all of which are technical (see Figure 1).

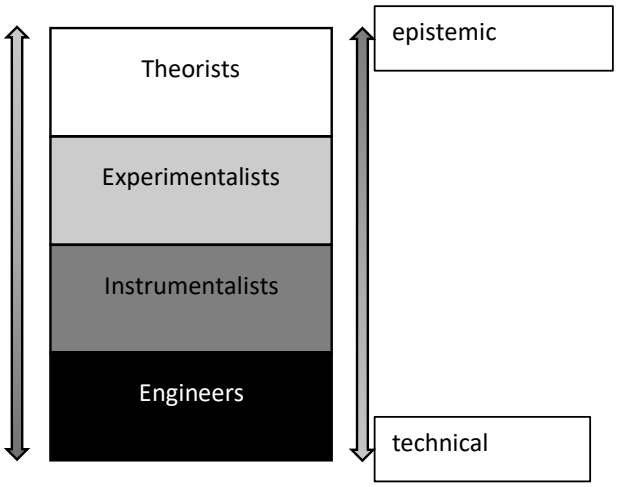


Figure 1. Epistemic hierarchy of communities in high-energy physics. Detector, as well as accelerator, researchers perform both epistemic and technical roles (which clearly differ, preserving the dichotomy), however, if the technical roles of the experimentalist (data analysis) are subjected to the measurement of the phenomena under scrutiny as an immediate goal, then even the epistemic goals of instrumentalists are resigned to improving the device designs (for example, of a particle accelerator), in connection with which the experimenters find themselves higher in the epistemic hierarchy. Here we assume the data-analyzing detector researchers to be experimentalists while both nondata-analyzing detector scientists and accelerator researchers are instrumentalists.

Engineering and epistemic justice in megascience

Thus, we have argued that the dissimilarity between engineering and other types of scientific activity do not resolve into the identity or constructive nature of the activity, but, first of all, are governed by perceived property relations and the rights claimed to experimental data that arise due to involvement in the data acquisition process. It is immixture in epistemically significant practices and affiliation in the data-harvesting community that come to the forefront in distinguishing an experimentalist from an engineer. Restrictions on the access of certain groups to epistemic practices based on their social or professional group membership and external identity raise the issue of epistemic justice.

Although epistemic injustice has been extensively discussed in the philosophical literature during the last decades (Fricker 2007; Anderson 2012; Medina 2012; Pohlhaus 2017), only very recently was the attention of scholars attracted to the internal workings of the scientific community (Grasswick 2017; Perović 2017; Pronskikh 2018; Pla-Julián 2018). The account by Fricker (2007) originally identifies two types of epistemic injustice and considers them in relation to prejudice against social identities of certain discourse participants. As was extensively discussed above, contemporary big science, especially high-energy physics, exhibits a complex social structure. The

community that constitutes it is stratified into subcommunities associated with certain epistemic and technical practices, which are unequal in their epistemic weight and value (Galison 1997). These communities develop various technical languages for communication, and some of them are functionally and linguistically subordinate to others which puts them in epistemically unequal positions (Pronskikh 2018).

In her original book, Fricker (2007, 158) suggested two types of epistemic injustice: testimonial injustice, which is due to systematic credibility discounting to people of certain disadvantaged or stigmatized social identities, and hermeneutical, which implies that members of socially marginalized groups lack resources to make meaning of their experience interpretable by the society. The remedies to both testimonial and hermeneutic injustices suggested by Fricker (2007) are respective virtues that individuals must exercise to counteract their prejudices. In Anderson (2012), the individual virtue-based remedies for epistemic injustices are challenged along two lines of argument. First, because cognitive biases are rooted deeply in the mind and have an automatic character, prejudicial hearers may discount the interlocutor's testimony because they perceive it incompetent or dishonest. Therefore, cognitive biases are difficult to control, although well-intentioned agents can train themselves to practice cognitive dissonance to discount their perceptions. Second, the credibility of the social groups can be discounted or favored not only on a transactional basis but also due to their belonging to a group (for example, certain ethnicity or using certain grammar). In the case, for example, of group favoritism or bias, there is no transactional injustice, on the contrary, in-group trust is vital in cases of division of labor (Anderson 2012, 170). Such cases cause, however, structural testimonial injustice and call for structural changes for their remedy. In her view, redesigning social institutions is unavoidable to mitigate structural epistemic injustice.

In the engineering context, the most relevant is seen as the concept of participatory injustice proposed by Hookway (2010) to clarify the forms of testimonial and hermeneutic injustices (Fricker 2007). Hookway (2010) pins down that "Participating is not just a matter of exchanging information: it involves asking questions, floating ideas, considering alternative possibilities, and so on." He concludes "epistemic injustice that is directed at someone's functioning as a participant in discussion, deliberation, and inquiry does not simply cause the victim to lose epistemic confidence more generally. Rather it questions the possession of capacities that are necessary for participation in these kinds of epistemic activities." (p. 6) Excluding engineers who, as we explained earlier, have most of the basic skills necessary for a scientist belonging to more epistemically significant discourses and practices, such as the collection and processing of experimental data, their discussion and presentation of the results of cognition outside, in our opinion is an example of epistemic injustice. In this regard, participatory injustice should be considered alongside other types of injustice, which is also structural in nature, i.e., requires institutional efforts, not just individual ones to transcend them.

Anatomizing the problem of institutional epistemic justice, Anderson (2012) points out that the epistemic segregation of the communities is just as unfair as ethnic or racial biases. However, collaborative learning and research can help overcome the bias of individuals and more privileged groups over less privileged ones. In the context of scientific and engineering communities, we believe that, in relation to scientific research and megascience in particular, collaborative learning and research can mean that representatives of separate communities (both detector and accelerator

communities and, within these communities, research and engineering) should not only complete the same training courses, but also jointly discuss and contribute to all stages of research, from accelerated particles and facilities to data analysis and phenomenological theoretical calculations. Moreover, mobility between communities must be ensured, providing the opportunity and ability to move from engineers to scientists. This will help both to transcend the perception of boundary objects as delimiters between epistemic and nonepistemic communities and to fulfill the ethical requirement of epistemic equality, which is considered a condition of epistemic democracy.

Conclusion

This paper examines the community structure in high-energy physics, which for decades has been considered as including instrumentalists, experimenters, and theorists. In our view, the first two communities are more correctly regarded as accelerator and detector researchers, which can be divided into several groups, including engineers and other specializations. I seek to address the noted issue of a shift in the identity of the experimentalists and explain it through the convergence of the constructive nature of work of the experimentalist with the engineer as well as the advent of specialists of a narrower profile in the place of the classical experimentalist with epistemic division of labor. Under the conditions of a similar nature of labor against the background of narratively fixed perceptions about their purported scientific expertise and horizons, this could entail a certain crisis of the experimentalist's identity. We note that identities of the experimentalist and engineer began to blur and overlap, and their activity formulas nowadays almost coincide. The basis of the external distinction between engineers and nonengineers, as before, is the orientation of their constructive activities either toward the artificial, technical nature (among the former) or by natural phenomena (among the latter). At the same time, the engineering, constructing nature of labor turns out to be characteristic of both scientists and engineers, and the formal orientation of the activity toward artificial nature as a functional role, as a rule, serves as the basis for the refusal of engineering specialists to participate in experiments and analyze data. I argue that the exclusion of engineers and other nonscientist specializations in megascience from epistemically most valuable discourses and practices was considered by us in the framework of the concept of participatory epistemic injustice. I suggest an avenue of approach to overcome participatory injustice, such as joint projects for engineering and nonengineering specializations, in which they cast themselves in epistemically equipollent roles.

References

Anderson, K., J. Boyett, et al. 2010. Understanding engineering work and identity: a cross-case analysis of engineers within six firms. *Engineering Studies*, 2 (3), 153-174.

Anderson, E. 2012. Epistemic justice as a virtue of social institutions, *Social Epistemology: A Journal of Knowledge, Culture and Policy*, 26, 2 (2012): 163–173, doi: 10.1080/02691728.2011.652211

Boon, M. 2011. In Defense of Engineering Sciences. On the Epistemological Relations between Science and Technology. *Techné. Research in Philosophy and Technology*. 15: 49-73.

Collins, H.M. & Evans, R.J. 2002. The third wave of science studies: Studies of expertise and experience. *Social Studies of Sciences* 32, no. 2: 235–296.

Fricker, M. 2007. *Epistemic Injustice. Power and the Ethics of Knowing* New York: Oxford University Press.

Galison, P. 1987. *How Experiments End*. Chicago and London: University of Chicago Press, 1987.

Galison, P. 1997. *Image and Logic: A Material Culture of Microphysics*. Chicago: University of Chicago Press, xxv+955 c.

Grasswick, H. 2017. Epistemic Injustice in Science In Ian James Kidd, José Medina, and Gaile Pohlhaus Jr. (eds.) *Routledge Handbook of Epistemic Injustice* pp. 13–26. Location: Routledge. doi:10.4324/9781315212043.

Hoddeson, L. et al. 2008. *Fermilab: Physics, the Frontier, and Megascience*. University of Chicago Press, Chicago, Illinois, 520 p.

Hookway, C. 2010. Some varieties of epistemic injustice: reflections on Fricker. *Episteme*, 7 no. 2: 151-163. doi:10.3366/E1742360010000882

Knorr-Cetina, K. 1999. *Epistemic cultures: how the sciences make knowledge*. Cambridge, Massachusetts: Harvard University Press, 1999. 352 p.

Latour, B. 1979. *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills: Sage, 272 p.

Martin, J. D. 2017. "Prestige Asymmetry in American Physics: Aspirations, Applications, and the Purloined Letter Effect." Science in Context 30(4):475–506.

Medina, J. 2012. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations* Oxford: Oxford University Press.

Murphy M. et al. 2015. Designing the Identities of Engineers. In Engineering Education and Practice in Context, Vol. 2. Springer.

Perović, S. 2018. Egalitarian paradise or factory drudgery? organizing knowledge production in high energy physics (HEP) laboratories. *Social Epistemology*, 32 no. 4: 241-261, doi: 10.1080/02691728.2018.1466933.

Pickering, A. 1984. *Constructing quarks: a sociological history of particle physics*. Chicago, Illinois: University of Chicago Press.

Pla-Julián, I. and Jose-Luis, D. 2018, Gender Equality Perceptions of Future Engineers, Engineering Studies, DOI: 10.1080/19378629.2018.1530242

Pohlhaus Jr., G. ed. 2017. Varieties of epistemic injustice, In *Routledge Handbook of Epistemic Injustice*, pp. 13–26. Location: Routledge. doi:10.4324/9781315212043.

Pronskikh, V. 2018. Linguistic privilege and justice: what can we learn from STEM? *Philosophical Papers*, 47, no. 1: 71-92, DOI: 10.1080/05568641.2018.1429739.

Traweek, S. 1988. *Beamtimes and Lifetimes: The World of High Energy Physics*. Harvard University Press, Cambridge, MA, 1988. 206 p.

Wenger, E. 1998. *Communities of Practice: Learning, Meaning and Identity*. Cambridge, UK: Cambridge University Press.

forthcoming in PSA2020/2021

The diversity–ability trade-off in scientific problem solving

Authors:

Samuli Reijula

Theoretical Philosophy, University of Helsinki

Jaakko Kuorikoski

Practical Philosophy, University of Helsinki

15 Jan 2021

The diversity-ability trade-off in scientific problem solving

Reijula, Samuli[1] and Kuorikoski, Jaakko[2]

[1]Theoretical Philosophy, University of Helsinki

[2]Practical Philosophy, University of Helsinki

15 Jan 2021

Abstract

According to the diversity-beats-ability theorem, groups of diverse problem solvers can outperform groups of high-ability problem solvers. We argue that the model introduced by Lu Hong and Scott Page (2004; see also Grim et al. 2019) is inadequate for exploring the trade-off between diversity and ability. This is because the model employs an impoverished implementation of the problem-solving task. We present a new version of the model which captures the role of 'ability' in a meaningful way, and use it to explore the trade-offs between diversity and ability in scientific problem solving.

Keywords— social epistemology of science; group problem solving; cognitive diversity; agent-based modeling; distributed cognition

1. Introduction

Modern science is a deeply collaborative enterprise. Most genuinely important intellectual challenges cannot be tackled by a single scientific discipline, let alone by individual researchers.

Science needs diversity – solving scientific research problems requires attaining specialized expertise and resources from a variety of perspectives.

Problem-solving groups in general are taken to benefit from diversity (Reagans and Zuckerman 2001; Mannix and Neale 2005; Jeppesen and Lakhani 2010; Steel et al. 2019). Among other important benefits, it is assumed that differences in how members of a group see a problem, in the cognitive resources they have at their disposal, and in the kind of heuristics they use, make it more likely that the the group as a whole has the resources to solve the problem. An important question, therefore, is whether the diversity of a group is in itself epistemically valuable, over and above the epistemic abilities of the group members.

Besides the empirical evidence cited above, a particularly influential argument in favor of diversity has been presented in the form of a mathematical theorem and an agent-based simulation. According to the *diversity-beats-ability (DAB) theorem*, groups of diverse problem solvers can outperform groups of high-ability problem solvers. This means that in assembling problem-solving teams, functional group diversity should sometimes be prioritized over selecting the most able individual members. Although they originate in computational social science, in management and organization studies, the DAB results have recently been also discussed in the philosophy of science (Grim et al. 2019; Singer 2019; Holman et al. 2018).

We argue that the "can" in the DAB theorem is ambiguous between several different modalities: in some of its uses, it is only a claim about conceptual possibility, whereas in its much advertised practical applications, it is clearly regarded as a more substantial possibility. This raises the question of when and under which exact conditions diversity really beats ability. We examine whether the original model by Hong and Page, and its further developments by Grim and associates, actually support the existence of the diversity-beats-ability phenomenon.

We show that due to their *impoverished task implementation*, these models cannot capture

interesting trade-offs between functional diversity and individual ability: the problem-solving tasks portrayed in the models are too difficult (i.e., random noise) for ability to make any difference to the outcomes. We develop a new version of the model with an improved problem-solving task. The new task representation allows our model to capture the role of individual ability in problem solving. Only when both diversity and ability really affect the outcome can the trade-off between them be studied.

We start by briefly presenting the DAB theorem and the associated simulation models, focusing on the latter. In Section 2, we highlight the "bait-and-switch" argumentative strategy used by Page to argue for DAB, showing that many of the modeling results supposed to support the theorem are problematic and do not replicate well. In Sections 3 and 4, we present our main argument: the model template used by Hong and Page as well as Grim and colleagues is ill-suited for exploring the trade-off between diversity and ability, because the problem-solving task is computationally implemented in a way that does not afford any advantage to individual ability or expertise. We introduce our version of the model, the stairway landscape, and demonstrate how it captures a substantial trade-off between diversity and ability. We draw two potentially interesting conclusions concerning the trade-off.

In this article, we are only concerned with the purely instrumental value of cognitive diversity; we are not arguing against the DAB phenomenon as such. We only ask whether the particular models we discuss are an informative and reliable way of exploring the possible trade-off, and provide what we regard as a better alternative way for doing so.

2. The diversity–ability trade-off in group problem solving

Consider design tasks such as designing an automobile, a space shuttle, or a piece of software, or scientific tasks such as measuring the mass of an elementary particle or discovering the structure of a macromolecule. Heterogeneous cognitive and material resources need to be applied to solve all these problems, and as the set of solution candidates is not known beforehand, a search for solutions is needed. Simon (1989) suggested viewing the scientific research process through the lens of heuristic search. For instance, scientists search for formulations of problems, experimental designs, patterns in data, mechanisms behind data, and implications of their theories. On some occasions, these multi-dimensional search trajectories result in beneficial epistemic design; in other cases, they yield research approaches of little cognitive value. Importantly, most scientific problems worth solving lie beyond the capacities of a single knower, and scientific progress relies on a successful division of labor and collaboration between researchers, research groups, and sometimes even between scientific disciplines. Hence, scientific research should be understood as a socially distributed problem-solving process.

Such a picture of collective search immediately suggests a possible trade-off. On the one hand, as Newell and Simon (1972) suggested, expert performance often relies on highly specific search heuristics. On the other hand, more diversity in the group's cognitive resources is beneficial, all other things being equal, as more varied resources provide access to larger portions of the solution space. Diversity may, however, conflict with individual ability. Experts are often more alike (in the relevant respects) than non-experts. Herein lies the trade-off: individual ability and group diversity both contribute to group performance, but, at least in some circumstances, the two factors may be in conflict.

Explicit modeling of the epistemic benefits of diversity in collective problem solving is

needed, because the phenomenon involves multiple group-level mechanisms as well as possible interactions between different epistemic, processual, and social factors. Therefore, purely verbal and conceptual theorizing is not a reliable tool for drawing out the implications of theoretical assumptions, and empirical (experimental or case-based) evidence does not usually unambiguously discriminate between alternative mechanistic explanations for why, in any particular case, diversity may or may not facilitate successful problem solving. Group problem solving has proved challenging to model, however. The computational implementation of the problem (task), cognitive resources (and differences therein), problem-solving behavior and cognition, and interaction between the group members all present difficult methodological and theoretical choices for the modeler, easily resulting in complex and intractable models with too many methodological degrees of freedom. Such models yield results which are hard to interpret. We believe that the heuristic-search paradigm proposed by Newell and Simon (1972) still provides the most promising approach for addressing these modeling challenges (see also Kauffman and Levin 1987; March 1991; Darden 1997). The models discussed and developed in this article join this tradition.

In a series of articles and books, Lu Hong and Scott Page have provided model-based evidence for the existence of the diversity-ability trade-off (Hong and Page 2001, 2004; Page 2008). They, in fact, use two distinct models to investigate diversity. The first model, introduced in Hong and Page (2001) and described in length by Page (2008) in the context of the diversity theorem, represents the problem to be solved as a binary string of finite length, where each bit could be seen as portraying a yes–no decision regarding a solution to a particular sub-problem (Kauffman and Levin 1987). A group of problem solvers of limited ability attempts to maximize a value function defined over the possible states of this string (potential solutions to the problem). Diversity is represented in the model by each agent having a different set of possible ways of

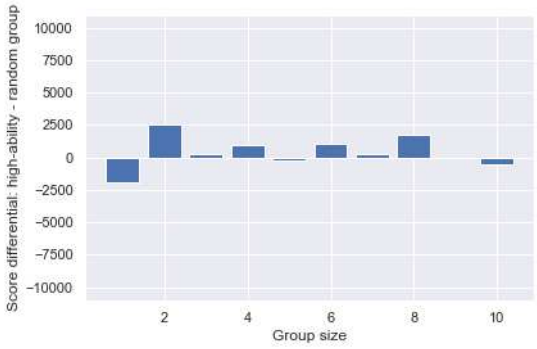


Figure 1: High-ability vs. random groups in the bit string model. The vertical axis represents the score differential between high-ability groups and random groups.

flipping the bits ("flipset heuristics") of the candidate solution string shared between the group members. Measures of problem difficulty can be assigned to alternative value functions (see Page 1996), and so the model can be used to represent a range of problems of different difficulty and complexity. This model template therefore corresponds well to pre-theoretic intuitions about how cognitive diversity can facilitate collective problem solving.

It is therefore rather surprising that the influential diversity-beats-ability results are not derived from this model. Our replication of the model in Hong and Page (2001) did not provide evidence to support the diversity-beats-ability phenomenon (see figure 1).[1] As the figure illustrates, no systematic difference emerges between groups of high-ability problem-solvers and groups of randomly selected problem-solvers. A more careful look at Page's 2008 argument reveals that it is based on evidence for the diversity theorem from an altogether different model introduced in Hong and Page (2004). We refer to this simplified model as the *ringworld model*. In sum, the substantial intuitions about diversity and ability in collective problem solving are first

1. For details about the bit string model, see Hong and Page 2001. All program code for the simulations and the generated data sets are available for download at https://osf.io/a6f5e/?view_only=fcee3f72db8643b9999ad19447f89886

formalized in one model, but the results are derived from a different model based on assumptions which do not correspond as neatly to the original intuitions. We find such a "bait-and-switch" argumentative strategy confusing, and not appropriate for transparent and epistemically sustainable use of theoretical models.

The argument in Hong and Page (2004) has a two-pronged structure: The basic assumptions of the ringworld model are used to derive an analytical proof intended to provide support for the theorem. However, as argued by Thompson (2014), the implications of the proof are unclear: even after technical corrections, the theorem only provides a highly abstract proof of possibility, and its implications for a non-technical interpretation of diversity are difficult to judge. Although we agree with Singer (2019) that the proof does rely on diversity and not merely on randomness (see Thompson 2014), it still remains the case that as such, the proof tells us little about the conditions under which the trade-off between diversity and ability can be expected to be significant. Mere logical possibility is not enough for the far-reaching practical implications suggested by Hong and Page. Their more persuasive evidence for DAB and its relevance for real-world group problem solving are derived from their agent-based simulation of the ringworld model. It is to this simulation that we now turn.

3. Problems in the Ringworld

The "computational experiment" used by Hong and Page to demonstrate DAB portrays a group of agents collectively searching for optimal solutions in a one-dimensional landscape. The discrete landscape consists of positions $1 \ldots n$ on the number line, wrapped as a circle.[2] Value function V defined over the set of positions assigns to each position a payoff value drawn from the uniform

2. It turns out that the circular topology of the landscape does not make a difference to the results, as the distance explored by the individual agents (and groups) typically does not exceed 20 steps along the 2000-step circle.

distribution [0,100]. The agents' goal is to find the largest possible values on this landscape. To do so, each agent employs a heuristic ϕ. A heuristic is defined as consisting of k different jumps of length $1 \ldots l$ (e.g., [1,5,11] and [3,4,12] are two examples of heuristics with parameters $k = 3, l = 12$). Starting from its current position, an agent sequentially applies these jumps along the landscape, and moves to a new position along the circle if the payoff associated with that position is strictly larger than the current one. When no further improvement is possible, the agent stops. The performance of an agent is defined as the expected payoff of the stopping points over the different starting positions of the landscape, and over a set of landscapes.

Hong and Page implement group problem solving behavior as sequential, iterative search. First, one agent initiates the search. As its local maximum is found, the second agent in the group takes the baton, and applies the jumps included in its heuristic as long as they lead to improvements. After all group members have taken their turn, a new round begins. The collective search stops when no agent can make further progress. Group performance is defined as the expected value of the position at which the group search stops.

In order to compare groups of high-ability problem solvers to more diverse ones, an exhaustive set of agents (with respect to possible heuristics) is first ranked according to their individual performance on a set of landscapes. A high-ability group of size g is constructed from the g highest performers in such a tournament, whereas the diverse group consists of g agents sampled randomly from the population.

In their model analysis, Hong and Page (2004) report results for various sets of parameter values. For example, for $l = 12, k = 3, n = 2000$ they find that that the best individual agents scored 87.3 whereas the worst agent's score was 84.3. For groups of 10, the high-ability group scored 92.56 and the random group 94.53. This difference in favor of the random group is the diversity effect discovered in the simulation. Similar results were found by Grim and associates

(2019), and we were also able to replicate the findings.

Hong and Page suggest that there are reasons to believe that the random group scored higher due to its diversity. An alternative way to express this finding is in terms of *effective group size*. In our replication, we noticed that the difference in performance ('performance differential') between the random and the high-ability group was strongly correlated (.65) with the difference in effective group size between the two groups, where effective group size was defined as the size of the group heuristic from which overlapping elements had been removed. In other words, the similarity between the members of a high-ability group results in the group being functionally smaller (from the perspective of the problem-solving task). As the performance of a group generally increases as its effective group size gets larger, it is not surprising that smaller effective group size leads to worse performance.

Going back to the original DAB theorem, however, the explanation above seems to capture only one side of the diversity–ability trade-off. Although the correlation between effective group size and performance is an indication of the functioning of the "diversity mechanism," it is still unclear why that effect is stronger than the influence of the "ability mechanism," i.e., the fact that some heuristics should lead to higher performance than others, and that those high-performing heuristics should be more common in high-ability groups. A closer inspection of the model provides a solution to this puzzle.

Unlike Hong and Page, we regard the effect sizes from the simulation as remarkably small, given that they originate from theoretical modeling where the modeler is free to explore a broad range of hypothetical scenarios. One would expect a purely theoretical model, purpose-built to examine and demonstrate a specific mechanism using heavy idealizations, to reveal relatively unambiguous effects of the modeled mechanisms. As a matter of methodological principle, we believe that conclusions drawn from agent-based modeling would be strengthened by showing

how the effect size can be manipulated by changing model parameters. In other words, being able to "turn the dials" and observe how changes in model inputs result in systematic changes in the modeled effect suggests that we have reached understanding about the dependencies between model inputs and outputs (see Woodward 2003; Aydinonat, Reijula, and Ylikoski 2020). Regarding the ringworld model, we argue there are two reasons to believe that the results reported by Hong and Page do not provide genuine insight into the diversity-ability trade-off.

First, with the parameter values studied by Hong and Page, in nearly half of the cases, the random group ends up with a *full heuristic*, that is, a heuristic consisting of all possible jumps $[1, \ldots, 12]$. Furthermore, only 13% of the random groups have an effective group size smaller than 11. Hence, even if the agents in the high-ability group can make the jumps leading to high performance, it is highly likely that the same jumps will also be included in the heuristic of the random group – there is simply no way the high-ability group could systematically outperform the random one.

Secondly, as Grim and his colleagues (2019) also noted, the purely random landscapes studied by Hong and Page are simply not hospitable to anything that could be meaningfully interpreted as “ability” or "expertise." For heuristic search to be applicable, the task needs to have some structure or redundancy that the heuristic can exploit (Kahneman and Klein 2009; Kauffman and Levin 1987). Hence, aggregated over several random landscapes, no significant performance differences emerge between the different heuristics. This is seen in the very small performance differences between the best and worst performing individual agents (see above) in Hong and Page’s simulations: the "ability mechanism" does not get any traction on the landscapes they studied. Therefore, we argue that the model does not appropriately capture the trade-off between diversity and ability.

Grim and his coauthors (2019) propose to remedy this problem by partially smoothing out the

random landscape (by adding interpolated values between randomly generated values). They argue that such a task representation can better capture ability, because on smoothed out landscapes individual performance is more transportable to other landscapes of similar smoothness. Yet a closer numerical examination of the results of this remedy again reveals only small differences between diversity and ability. Even on smoothed random landscapes, the expected performance difference between best performing and random individuals is minute. This suggests that these landscapes simply do not represent a problem that is suitably complex for exploring trade-offs between ability and diversity.

4. Modeling the diversity–ability trade-off on stairway landscapes

In order to better understand the tension between diversity and ability, we need to portray scenarios where also ability plays a role. In our own simulations, we introduce a type of problem where high ability – either at individual or group level – leads to noticeably increased performance. In science, having the right methodology for the problem at hand often sharply increases the epistemic payoff. Our stairway model differs from the Hong and Page ringworld model only in problem structure. The specifications of agent and group behavior remain the same as in the ringworld model. In generating problem landscapes, we start from the uniform noise distribution employed by Hong and Page. On top of those landscapes, however, we superimpose an increasing sequence of values, where the positions of the values are separated by intervals drawn from a finite set of integers in $1 \ldots l$ (see figure 2). We call this set the *step set*.

For an agent to climb the increasing subsequence, the *stairway sequence*, it must possess the

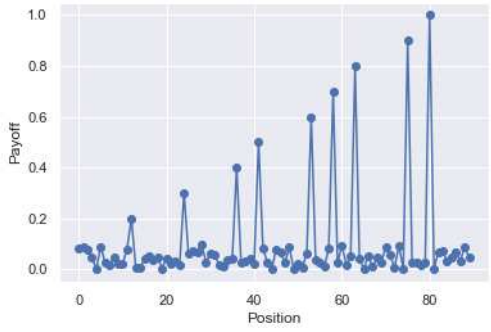


Figure 2: A stairway landscape with step set $\{5, 12\}$, and, therefore, step set size 2.

heuristic jumps corresponding to the steps used to generate the sequence (e.g., $[5, 12]$ in figure 2). This strongly favors some heuristics over others: whereas an agent who does not possess the full step set is bound to remain in the noise region of the landscape, a "high-ability" agent that has the necessary heuristic can climb through the whole sequence (and even reach the maximum payoff on the landscape, normalized to 1.0).

Figure 3 illustrates outcomes from our model with parameters values corresponding to those studied by Hong and Page (2004) and by Grim an his colleagues (2019). The left panel presents the difference between the performance of high-ability and random groups (positive values standing for high-ability group advantage, and negative values, for random group advantage). The results indicate that with these parameter values, stairway landscapes always favor high-ability groups. Especially when the group size is small, because it is made up of high-performing individuals (who typically possess valuable elements of the step set) the high-ability group performs significantly better than the random group. The right panel presents the difference between the redundancy of heuristics between the high-ability and random group (value 0 means that the overlap of heuristics in both groups is the same). As suggested by findings by Hong and

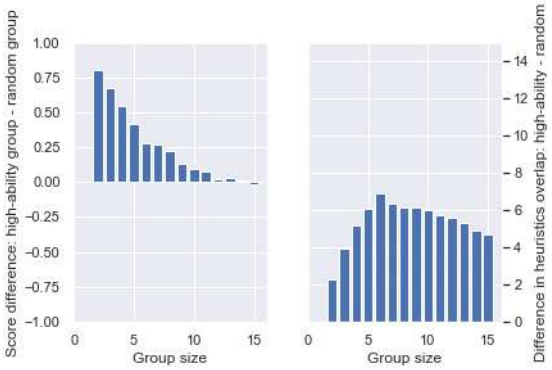


Figure 3: High-ability vs. random groups on a stairway landscape, step size 3. ($k = 3, l = 12, n = 2000$; 100 repetitions over 100 landscapes)

Page (2004), random groups tend to have comparatively lower levels of overlap in their heuristics. As group size increases, the redundancy in the high-ability group increases more than in the random group. This suggests that when the group size is larger, random groups again begin to approach the full heuristic, which obviously is sufficient for climbing the stairway sequence. For this reason, at group sizes larger than 10, random groups catch up, and no significant performance difference is observed between high-ability and random groups (left panel).

We argue that this tension between the "ability mechanism" and the "diversity mechanism" captures the trade-off addressed by the DAB theorem. What happens, however, when the level of ability or expertise required by the task changes? Different levels of task difficulty can be represented by stairway landscapes with different step set sizes. For example, landscapes with step set sizes up to three lie within the abilities of the individual agents studied in the simulation ($k = 3$). Climbing the stairway for step sizes larger than 3 requires pooling heuristics from several agents.

Figure 4 summarizes tentative findings from our studies with landscapes of varying difficulty. In the figure, group size is represented on the horizontal axis, and step set size (complexity of the

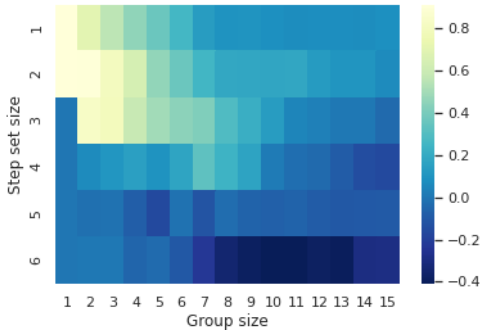


Figure 4: High-ability-vs-random group performance differential on stairway landscapes (50 repetitions, each over 50 landscapes).

problem) on the vertical axis. The color represents the performance differential between the high-ability group and the random group; lighter shades standing for high-ability group advantage. A genuine trade-off between diversity and ability can be seen. Observe the contrast between the upper-left quadrant, where ability dominates, and the lower-right, where random groups have a slight advantage over the high-ability groups; ability dominates when group size and step set size are small, whereas diversity leads to better performance when the group size and step set size are larger.

Finally, our results suggest a conceptual distinction between the complexity and difficulty of a problem: perhaps not surprisingly, ability dominates when the problem is simple in the specific sense that multiple cognitive resources do not need to be combined to solve it. Note that if the problem is simple in this sense, this does not necessarily mean that it is easy to solve. When the problem becomes complex, requiring efficient division of cognitive labor, the diversity effect begins to dominate over individual abilities. The results demonstrate how diversity and group size begin to outdo individual ability only when the problem complexity exceeds the cognitive resources of any single individual.

One could object to our stairway model on seemingly similar grounds to the ones on which we based our criticism of the original ringworld model. We questioned the DAB results on the basis that the model was built to favor diversity over ability. Why would our model fare any better, as it was clearly built to favor ability over diversity? This objection misses our point, however. Our argument is that the original model cannot be used to model the *trade-off* between diversity and ability, because it cannot be used to represent the gains from ability. Of course, we fully admit that the stairway landscape is built to favor ability, but the model nevertheless also retains the gains from diversity. Stairway landscapes give both ability and diversity their due, and, therefore, can illuminate the trade-off between them. This, we argue, was the original and interesting interpretation of the DAB results to begin with.

5. Conclusions

The original results by Hong and Page do not provide reliable evidence for the diversity-beats-ability theorem because the ringworld model, especially its task implementation, does not allow for ability to adequately influence individual or group performance. This one-sidedness implies that their model cannot be used to explore the possible trade-offs between diversity and ability in problem-solving groups. Our exploration of stairway landscapes illustrates how the results by Hong and Page (2004) rely on a problematic task structure to get their results.

Stairway landscapes provide a better model for "medium-hard" problems which require specialized abilities and true division of cognitive labor. Such landscapes can be used to model the interplay between diversity and ability relevant, and its effects on the division of cognitive labor in science.

Our tentative modeling results suggest a trade-off between diversity and ability. Ability is

favored when the problem is moderately difficult, requiring only a few different expert heuristics, and when groups are small. Diversity is favored when the problem is complex, requiring multiple component solutions, and when the groups are large. A further qualitative effect can be observed at the point where problem complexity increases beyond the capacity of a single agent and necessitates division of cognitive labor: simple problems solvable by individuals favor ability regardless of group size.

Acknowledgments

We thank Kristina Rolin, Inkeri Koskinen, Renne Pesonen, and the other members of the TINT group (University of Helsinki), as well as the participants of Diversity in Science workshop (Tampere University, 5 May 2019) and the poster sessions at EPSA 2019 (University of Geneva) for their helpful comments. Thanks to Kate Sotejeff-Wilson for editing the manuscript. This research was carried out as a part of the project "Social and Cognitive Diversity in Science" funded by the Academy of Finland.

References

Aydinonat, N. E., S. Reijula, and P. K. Ylikoski. 2020. "Argumentative landscapes: the function of models in social epistemology." *Synthese,* forthcoming.

Darden, L. 1997. "Recent work in computational scientific discovery." In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society,* 161–166. Mahwah, New Jersey: Lawrence Erlbaum.

Grim, P., D. J. Singer, A. Bramson, B. Holman, S. McGeehan, and W. J. Berger. 2019. "Diversity, ability, and expertise in epistemic communities." *Philosophy of Science* 86 (1): 98–123.

Holman, B., W. J. Berger, D. J. Singer, P. Grim, and A. Bramson. 2018. "Diversity and democracy: Agent-based modeling in political philosophy" [in en]. *Historical Social Research* 43 (1): 259–284.

Hong, L., and S. E. Page. 2004. "Groups of diverse problem solvers can outperform groups of high-ability problem solvers" [in en]. *Proceedings of the National Academy of Sciences of the United States of America* 101 (46): 16385–16389.

———. 2001. "Problem solving by heterogeneous agents." *Journal of economic theory* 97 (1): 123–163.

Jeppesen, L. B., and K. R. Lakhani. 2010. "Marginality and problem-solving effectiveness in broadcast search." *Organization science* 21 (5): 1016–1033.

Kahneman, D., and G. Klein. 2009. "Conditions for intuitive expertise: A failure to disagree." *American Psychologist* 64 (6): 515–526.

Kauffman, S., and S. Levin. 1987. "Towards a general theory of adaptive walks on rugged landscapes." *Journal of Theoretical Biology* 128 (1): 11–45.

Mannix, E., and M. A. Neale. 2005. "What differences make a difference? The promise and reality of diverse teams in organizations." *Psychological science in the public interest* 6 (2): 31–55.

March, J. G. 1991. "Exploration and exploitation in organizational learning." *Organization science* 2 (1): 71–87.

Newell, A., and H. A. Simon. 1972. *Human problem solving.* Vol. 104. 9. Prentice-Hall Englewood Cliffs, NJ.

Page, S. E. 2008. *The difference: how the power of diversity creates better groups, firms, schools, and societies.* Paperback. Princeton, N.J. ;Woodstock: Princeton University Press,

———. 1996. "Two measures of difficulty." *Economic Theory* 8 (2): 321–346.

Reagans, R., and E. W. Zuckerman. 2001. "Networks, diversity, and productivity: The social capital of corporate R&D teams." *Organization science* 12 (4): 502–517.

Simon, H. A. 1989. "The scientist as problem solver." *Complex information processing: The impact of Herbert A. Simon,* 375–398.

Singer, D. J. 2019. "Diversity, not randomness, trumps ability." *Philosophy of Science* 86 (1): 178–191.

Steel, D., S. Fazelpour, B. Crewe, and K. Gillette. 2019. "Information elaboration and epistemic effects of diversity." *Synthese,* forthcoming.

Thompson, A. 2014. "Does diversity trump ability?" *Notices of the AMS* 61 (9): 1–024.

Woodward, J. 2003. *Making things happen : A theory of causal explanation.* Oxford studies in philosophy of science. New York: Oxford University Press.

Learning From the Shape of Data

Abstract

To make sense of large data sets, we often look for patterns in how data points are "shaped" in the space of possible measurement outcomes. The emerging field of topological data analysis (TDA) offers a toolkit for formalizing the process of identifying such shapes. This paper aims to discover why and how the resulting analysis should be understood as reflecting significant features of the systems that generated the data. I argue that a particular feature of TDA—its functoriality—is what enables TDA to translate visual intuitions about structure in data into precise, computationally tractable descriptions of real-world systems.

1 Introduction

"Learning from the shape of data" describes an expansive portion of scientific activity. One common example is curve-fitting, in which a data set is visualized on a two dimensional grid, and we infer that the underlying mechanism generating the data can be characterized by a function with a similarly shaped plot.

As new techniques are developed to gather, store, and analyze large quantities of high-dimensional information, its increasingly difficult to visually identify and interpret relevant shapes. While we can scale up familiar curve-fitting tools, such as linear regression, we know there is more structure to be harnessed from large data sets than these methods can reveal.

One relatively new method of identifying "shapes" in data sets is topological data analysis (TDA). Topology is the study of the properties of shapes that are invariant under continuous deformations, such as stretching, twisting, bending, or re-scaling. TDA aims to identify the essential "structure" of a data set as it "appears" in an abstract space of measurement outcomes.

The simplest application of TDA is a type of *cluster analysis*–a method of identify "clusters" of data points that are "more similar" to one another than the wider body of data. While this is relatively conducive to interpretation (as revealed "groupings" in the system being analyzed), TDA can also identify more complex shapes including "holes", "voids", and "tendrils" with no intuitive interpretation.

This paper is an investigation into why and how the resulting analysis should be understood as reflecting significant features of the systems that generated the data. In particular, I will argue that the relevance and utility of TDA stems from a particular feature: the *functoriality* of the relationship between the shapes it picks out and their symbolic representations.

In section 2 I describe TDA in detail. Section 3 explains what functoriality means and how it justifies the use of TDA despite interpretational challenges. In section 4, I relate this discussion to philosophical work on the contents of and relationships among physical theories. Section 5 examines the role of spatial reasoning in TDA, and how its functoriality enables integrating this informal activity into a formal data analytic framework.

2 Topological data analysis

The phrase "topological data analysis" is used to refer to a variety of data science practices that use tools from algebraic topology to make inferences about the "shape" of data clouds as they appear in the "space" of possible observations. Here, the term *data* refers to a set of real vectors corresponding to a series of observations. This is an adequate definition for capturing natural language use of the term, but one might object that it does not necessarily capture what data *is*. One of the goals of TDA is to circumvent some of the arbitrariness involved in presenting data as real vectors. A *data cloud* can thus be thought of as a visual representation of this set of vectors as "points" in a (high dimensional generalization of) space. The abstract "space" where data lives is generally some form of *metric space*, or set X of points (including at least the data points) together with a notion of "distance" $d(\ ,\ )$ between the points. For example, I may have data about the weights of a collection of potatoes. The distance between these data points would just be the pairwise difference in weight between two potatoes according to a fixed unit, e.g. pounds.

A characteristic problem of analyzing large data sets is deciding how to

combine many different types of measurements into a shared metric space. I can also add information about the length, color, number of eyes, etc. for each potato, creating an n-dimensional space, where n is the number of potato attributes. The "distance" between two data points is now some combination of the distances given by weights, lengths, color, etc. But how should the notions of distance given by each variable combine into "distance" in the total space of possible variable values? The standard way of aggregating one-dimensional metrics into a shared metric space is to imagine each metric as an axis in an n-dimensional Cartesian grid, with distance given by the Cartesian distance as follows. Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ be two sets of potato measurements. Then $d(x, y) = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2}$. Setting aside the fact that there are other viable options for constructing distances from these values, notice that this expression does not include units. Should weight be presented in pounds or tons? Of course we know how to translate between these two units, and we consider the choice more of notational convenience than theoretically meaningful. But if we are looking to the "shape" of data for information about the system being measured, the data cloud will look much more "flat" if we use tons rather than pounds. It is thus desirable to consider properties of the data cloud that do not depend on the particular choice of metric space or unit, but which are shared by a variety of plausible modeling choices.

Such considerations motivate the use of *topological*, as opposed to geometric methods. Topology is the mathematical field that studies properties of shapes that remain constant under stretching, twisting, or otherwise deforming. Topologists attend to more general features of metric spaces that would be present under different modeling assumptions, called *topological invariants*. Since data sets are finite, although they may suggest some underlying shape, they likely will not do so uniquely. This is the standard curve-fitting problem in higher dimensions: for any discrete set of points, there are an infinite number of continuous curves (or shapes) that contain (or approximate) the locations of those points. As with the curve-fitting problem, external considerations guide the choice of continuous object, rather than just the bare, uninterpreted set of data points. One may have a priori reasons to expect that the "right" curve is quadratic, for example.

2.1 Clusters

The simplest example of TDA, and the one most broadly used by data scientists generally, is a type of cluster analysis. The idea behind cluster analysis is to ask: do my data points naturally divide into sub-categories of data points more similar to one another than the overall space? Such a situation indicates that there is some non-trivial structure underlying the data associated with such groupings, which one may interpret as "natural kinds" in the space. Cluster analysis is in this way closely related to regression analysis—clusters point towards a correlation among variables, one of the main "signals" data scientists hope to read off of large data sets. For example, biological species are sometimes individuated as "homeostatic property clusters" of organisms that are stably more similar to one another than to other organisms (Boyd, 1999).

In scientific contexts, external considerations about the type of data under consideration tends to influence how one chooses to carve a data set into clusters. For example, only features considered relevant to fitness will likely factor into the the similarity notion that underlies species clustering. Moreover, traditional clustering algorithms such as k-means will require a pre-specification of the number of clusters to be identified, which will likely come from preconceived notions of the expected number of groupings. For example, a clustering of voter data might pre-suppose that voters will split into two clusters along partisan lines.

Even in the absence of such guidance, natural clusters may be easily "seen" when the data is graphed. With larger and higher dimensional data sets to analyze, these heuristics are less useful, and data scientists would prefer a principled algorithmic approach to clustering. This would amount to a function that takes metric spaces (X, d)—here understood as data sets $X = \{x_1, ..., x_n\}$ with a notion of "distance" $d(x_i, x_j)$—as inputs, and outputs *partitions* of that data into clusters of data points that are "close together."

2.2 Constructing Shapes

The most common method to construct a shape from a data cloud is roughly as follows. Enclose each data point in a "ball" of radius ε centered on that point. As ε gets larger, the cloud will cease to look like isolated points and start to gain shape. Once it gets too large, though, we are left with a single shapeless blob. We use this idea to construct a *simplicial complex*, beginning

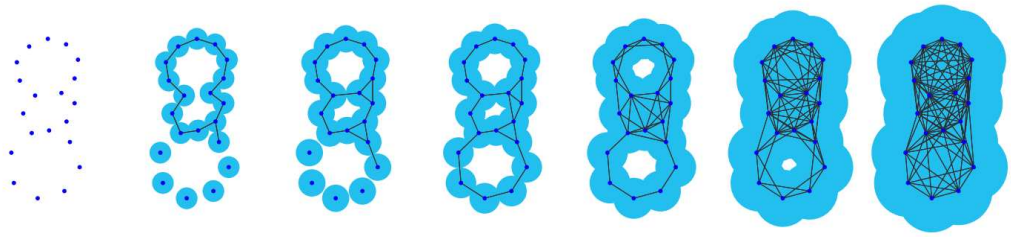


Figure 1: Constructing a Čech complex as ε increases, from Bubenik (2015).

with the data points as vertices.[1] Where 2 balls intersect, we add an *edge* between them. When 3 balls intersect, we add a *face* enclosed by the three edges. This process continues, creating higher dimensional *n-faces* where $n+1$ balls intersect. The result is called a *Čech complex*.[2]

This is an intuitively plausible way to construct a discrete shape from a data cloud. A clustering can be "read off" of a Čech complex by grouping data points according to whether they are connected in a single component of the complex. This may be complicated by the presence of noise—a single anomalous data point might connect otherwise robustly distinct clusters. This can be side-stepped by either looking at only regions that are highly connected, or avoided altogether by filtering and "cleaning" the data prior to analysis.

2.3 Holes and voids

Identifying the clusters of a simplicial complex appears is a special case of a more general phenomenon of *homology*. Homology is a method of classifying shapes by looking at how many "holes" the shape has. No matter how much you stretch and twist it, a circle will always have a "hole" in it, a sphere will always have a *void* or *cavity*, an innertube will always have the "donut hole" as well as a void in the interior that inflates.

When we look at the connected components of a Čech complex, we are considering the H_0-*homology* of the complex (considered as a topological space). We can similarly attend to the H_1-*homology* of the complex by

[1] See Hatcher (2002) section 2.1 for a precise definition of a simplicial complex.

[2] In practice, TDA employs a more computationally tractable approximation thereof, called a *witness complex*. See Carlsson (2009) section 2 for details.

looking for "holes," or the H_2-*homology* by looking at "cells," and so on to higher dimensions with less intuitive interpretations.

Example 1 (Cosmology)**.** van de Weygaert et al. (2011) study the homology of density level sets of an ensemble of randomly generated cosmic mass distributions. They analyze the evolution of H_1, H_2, and H_3-homology over time in n-body simulations, revealing characteristic patterns of different dark energy models. They show how homology can track cosmological structures of independent interest to physicists, such as matter power spectra and non-Gaussianity in the primordial density field.

2.4 Persistence

The motivating idea behind the construction of a Čech complex is that we can imagine data as being uniformly sampled (with noise) from some underlying "shape" in the metric state space, and we can use these data points to infer the global structure of the "object" we are sampling from. The more samples we look at, the more accurate our picture of the shape will be. For sufficiently small ε-balls, the complex will not have any more structure than the bare data set. Similarly, when the balls get too large, there is nothing more to look at than a giant blob. The "right" choice of ε is at some intermediate size, but how should it be chosen? If we chose an ε that is too small, we will get a shape with a lot more holes, disconnected components, etc., than we think are meaningful. In other words, we retain some of the noisy features of the data cloud that we were trying to eliminate. But we risk going to far, and making ε large enough to obscure both noise *and* meaningful information from the data.

A natural way to solve this problem is to look at many different choices of ε, and use external considerations to decide which gives the best resolution of the data shape. Two more problems arise when we do this, though. For one, the whole point of data analysis is to simplify and compress information about a system, and having a variety of different models we can choose from does not simplify matters. Second, there may be different features that arise at different resolutions that are equally significant, and this multi-level picture can get lost if we have to choose a single model among the many possibilities. For example, data may be dense in some regions but sparse in others, where relevant shapes require larger ε-balls to be "seen".

The key insight that unlocked the power of TDA was the idea of "topolog-

ical persistence," introduced to data analysis in (Edelsbrunner et al., 2002). Briefly: instead of picking a particular resolution to look at, we look at them all, but take advantage of a trick from algebraic topology to connect complexes at different scales in a sophisticated and efficient way The result is the association of a data cloud with a *persistence module* that encodes how the cloud changes structurally as ε increases. Homology is then computed for these modules, and the result is typically expressed as a *homological barcode*, as in figure 2. The "bars" begin when a feature is "born" and end when it "dies." Short intervals in barcodes are often attributed to either measurement noise or inadequate sampling, whereas long, "persistent" bars are thought to reveal real geometric features of the space being sampled from.

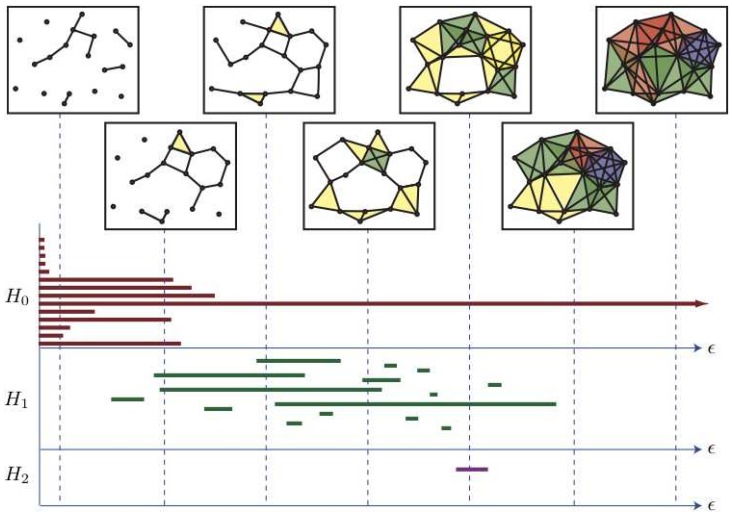


Figure 2: Example of a homological barcode, from Ghrist (2008).

Not only is this decomposition more computationally tractable to analyze than (sets of) complexes, but the barcode itself provides a visual summary of behavior as ε increases. When the number of features is large, data analysts will also sometime use *persistence diagrams* instead of barcodes.

2.5 Stability

One way to interpret ε is as a modeling parameter, corresponding to the resolution or scale we use to construct a shape from the data cloud. The persistent features of a Čech complex are those that are *stable*, or robust under

perturbations of the parameter value. Longer bars in barcodes represent features that appear for a wider range of ε values, indicating that these features are robust and unlikely to constitute mere noise. Cohen-Steiner et al. (2007) made this precise by proving that for a large class of constructions (including Čech complexes), persistence diagrams are *stable*, meaning that small perturbations of the initial data set result in correspondingly small changes in the resulting persistence diagram.

We can use this same method to consider stability across other indexing parameters as well at fixed resolution, as in the following example.

Example 2 (Arteries)**.** Bendich et al. (2016) employ topological data analysis to study the structure of arteries in the human brain. They uniformly sample a large number of points from a blood vessel diagram (weighted by thickness of vessel), and construct a Čech complex from this data cloud, analyzing the H_0 and H_1 persistence diagrams over the growing size of ε-balls in the Čech complex. They look at persistent H_0 over a stack of "horizontal slices" of the artery diagram.

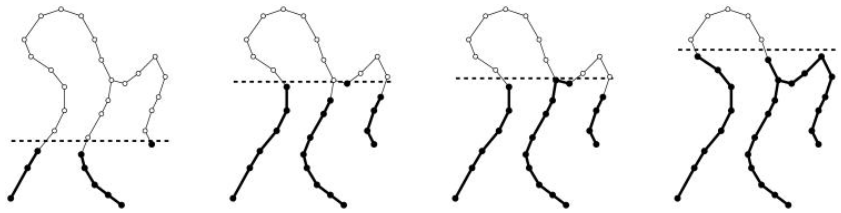


Figure 3: Horizontal slices of the artery diagram, from Bendich et al. (2016).

The authors found significant correlation between certain features of these homological barcodes and the age and sex of the subjects, with the age correlation a significant improvement over previous attempts at analyzing similar data. For example, older brains tended to have the longest bars in the latter barcodes.

In this example, persistence is indexed over the parameter of height. One can also analyze persistence of homological features over time.

Example 3 (Time-series data)**.** (Perea and Harer, 2015) demonstrate that persistent H_1-homology over time can be used to detect periodicity in time-series data by embedding it into a higher dimensional space. Note that in the absence of such an embedding, time series data displays no "loops" (since

prior points in time are never revisited), so as it stands, it is not conducive to analysis of homology. It is fairly common for data analysts to modify their data to match their methods in this way, rather than the other way around.

We can thus understand persistence modules as assembling a sequence of $(n-1)$-dimensional models indexed by an n^{th} parameter, such as resolution or time. Dimensionality reduction is a common feature of data analysis techniques. Data often comes in the form of large vectors, and the goal is often to *compress* them—express as much of the original information as possible with in as few dimensions as possible. This amounts to selecting features or parameters of interest and suppressing the rest in order to highlight general patterns. Reducing data models to 2-3 dimensions also makes them more visualizable, making them more useful to researchers to observe patterns, as well as easier to communicate to the public. Persistence modules provide the benefits of low dimensional visualizability without throwing away the information in the extra dimensions.

3 Functoriality

Most practitioners will admit that the interpretation of homology in data is unclear. While increasing in popularity of late, TDA (beyond mere cluster analysis) is still relatively niche. It is often reserved for situations in which traditional data analysis tools have failed to bear fruit, and TDA is one of many attempts to gain insight into the data.

Data scientists rarely feel the need to justify their use of TDA beyond the fact that it seemed to pick up on a relevant pattern in a particular situation. But when pressed, or in more comprehensive theoretical contexts, the use of TDA is usually explained by the fact that homology has a particularly nice property that makes it a reliable data analysis tool: *functoriality*.

To understand this, we'll need to look a bit deeper into how TDA functions. TDA summarizes the shape of a Čech complex built from a data cloud in terms of a *homology group* $H_n(X)$. For each group, $H_n(X)$ essentially characterizes how many "holes" are present in each dimension. This makes it easy to describe the shape computationally, as groups are more easily described symbolically than shapes. But in order for this symbolic representation to to be useful, we need to be able to identify which "holes" in our complex correspond to which symbolic representation, and we need to be able to track the holes as we evolve the complexes. We can do this,

because homology is *functorial* in the sense that more than just translating complexes to groups, it tells us how to translate maps between complexes into maps between groups while preserving all relevant topological information.

The functoriality of homology enables us to do three important things, which are essential to its utility in analyzing data: identify local structures, connect complexes as parameters vary, and compare complexes constructed from different samples. We can identify local structures via inclusion maps that pick out particular clusters, holes, and voids. We can then evolve these complexes by varying parameters of interest, and see which features persist. Lastly, we can perform an additional robustness check on our results by comparing clusters generated with different sub-samples of our data, in a way analogous to bootstrapping in statistics (Chazal et al., 2015).

Thus data scientists study persistent homology, not because they think of "counting holes" as the right way to characterize data, but rather because TDA has a particular feature–functoriality–that make it a reliable tool to use. Since persistent homology has this nice property, data scientists will often shoe-horn questions about data into the shape of a homology problem in order to make it tractable. For example, they might add extra edges to a Čech complex to turn open chains into closed loops. Or they might chose a particular dimensional reduction in which loops arise, as in Perea and Harer (2015).

One can also modify TDA to examine how clusters are shaped. For example, "tendrils" emanating from the core of a cluster can be tracked via the persistent H_0-homology of the resulting data cloud once that core is removed. Nicolau et al. (2011) use this technique to classify breast cancer types.

Figure 4: Visualization of data featuring tendrils.

While the recent proliferation of these methods might be dismissed as mere hammer-nailing, it should rather be said that since we have very few

tools to work with, we had better hope this problem can become nail-shaped.

If I am correct about the significance of TDA's functoriality, then we should expect that other fruitful data analytic methods can be understood functorially. Indeed, Bubenik and Scott (2014) express persistent homology as a special case of a more general kind of functor, and Carlsson and Mémoli (2013) demonstrate how a functorial account of clustering algorithms (including H_0 persistent homology) provides conceptual clarity.

4 Category Theory

The role of functoriality in justifying the use of TDA is suggestive of recent literature in the philosophy of physics advocating for a functorial account of intertheoretic relations. This literature is inspired by Halvorson (2013), who argues that one should understand the content of a scientific theory as a *category* of models of that theory. That is, as a collection of theoretical models, plus relationships (structure preserving functions) between the models. On this view, the appropriate way to understand relationships between theories is using a *functor*—a map that takes models to models and relations to relations in a consistent way. Once framed in this way, philosophers can use tools from *category theory* to enrich their understanding of these theories and how they relate to one another (Weatherall, 2017; Rosenstock, 2019)

We can conceive of TDA as a special case of this general category theoretic framework for characterizing scientific theories, or as a prefer to think of them, *representational frameworks*. We begin with a "metric space" representational framework for our empirical data. This consists of (finite) metric spaces, along with relationships between metric spaces (isometries, embeddings, etc.), forming category **FinMet**. We also have a "topological" representational framework of "shapes" that our data might have, and structure preserving maps between them forming a category **Comp** of simplicial complexes. And we have an algebraic category, **HomAlg**, of homological algebras.

In this language, we articulate a "reading" of shapes from a data set as a functor $F : \mathbf{FinMet} \to \mathbf{Comp}$, such as the functor F_δ that takes a metric space its Cěch complex of radius δ. And we can transform this topological framing into an algebraic framing via a functor from **Comp** to **HomAlg** (the "homology" functor). And we can construct a category **PDiag** of persistence diagrams, associated with our underlying data model again by a functor from

FinMet to **PDiag**.

There are lessons to be learned from this relationship between TDA and this philosophy of physics literature in both directions. Philosophers benefit from a fruitful example outside of physics, and one that incorporates many "levels" of abstraction from initial data to more abstract representations. Conversely, formal philosophical work can help elaborate the sense in which theoretical content is "preserved" in these functorial transformations. In particular, Rosenstock (2021) illustrates how reflection on the structure of a data set influences and constrains the ways in which it can be clustered.

5 Spatial inference

The goal of data analysis is to identify patterns in data that provide concise, comprehensible summaries of the system that point towards features of significance in broad classes of systems. Such recognition of patterns of sufficient generality without overfitting is the holy grail of artificial intelligence and machine learning research. In the meantime, scientists rely heavily on visual intuition to guide inquiry, experimenting with parameters and data filtering until it "looks right".

TDA removes *some* of the arbitrariness of this process by enforcing a consistent methodology to the identification of patterns once these discretionary setup choices are made. But intuitions are not abandoned entirely at this stage, since the resulting analysis still has to fit with preconceived notions of natural categories and interesting patterns in order to be of interest to practitioners. Patterns found through random applications of TDA might lead scientists to look for corresponding features of interest in a system, but if these cannot be found, the shapes identified in the data remain merely curiosities. In example 2, if barcodes did not track gender and age but some other feature that we do not independently classify as a natural kind, they would likely be omitted from the published analysis.

The difficulty of interpreting higher dimensional homology thus requires extensive human discretion to be empirically useful. As TDA is a second-line resource for data that is particularly intractable to analyze, which puts creativity at the center of its application. We might wonder whether such an informal process of intuitive speculation about the shape of data can be incorporated into a formal epistemic story about the structure of topological data models. Here, we can learn much from the vast literature on diagram-

matic reasoning in Euclidean geometry. Critics of the rigor of reasoning from diagrams in geometric 'proofs' point to the fact that such proofs use a particular illustration to make an inference about all possible illustrations. However, philosophers of mathematical practice have recently come to appreciate the role of diagrams in generating and communicating geometric knowledge. Manders (2008) argues that ancient geometers were careful to rely on diagrams only for demonstrations about what he calls *co-exact* features—those that are relatively insensitive to the range of variation in possible visual representations, such as part-whole and boundary-interior relationships (and of course, homology). Mumma (2010) takes this a step further and develops a formal account of Euclidean proofs that includes both sentential and diagrammatic components.

Similarly, data analysts are concerned with ensuring that inferences about data rely only on real structural features of observations, rather than incidental features of how data visualized. At issue is the level of generality one can adopt when making inferences from a single visual representation of data, picked somewhat arbitrarily from an ensemble of possible alternative, equally valid representations. TDA resolves this issue by requiring that the analyzed features of data models be *functorial* with respect to maps that preserve what they take to be the relevant structural features of models, and *persistent* across parameters when the "right" value is not known.

6 Conclusion

This paper argues that the functoriality of homology is critical to TDA's utility in revealing and interpreting structural features of data sets. In brief, topological features of data sets are visually salient to humans and aid in our reasoning in understanding. The functoriality of persistent homology ensures that reasons we had for thinking topological features were meaningful are preserved in the translation from data cloud to homological barcode, while enabling various robustness tests on the resulting analyses. There are promising future directions for exploring the relationship between topological data analysis and recent philosophical work on the content of and relationships among physical theories.

References

Bendich, P., J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer (2016). Persistent Homology Analysis of Brain Artery Trees. *The Annals of Applied Statistics 10*(1), 198–218.

Boyd, R. (1999). Homeostasis, species, and higher taxa. *Species: New interdisciplinary essays 141*, 185.

Bubenik, P. (2015). Statistical Topological Data Analysis Using Persistence Landscapes. *Journal of Machine Learning Research 16*, 77–102.

Bubenik, P. and J. A. Scott (2014). Categorification of Persistent Homology. *Discrete & Computational Geometry 51*(3), 600–627.

Carlsson, G. (2009). Topology and Data. *Bulletin of the American Mathematical Society 46*(2), 255–308.

Carlsson, G. and F. Mémoli (2013). Classifying Clustering Schemes. *Foundations of Computational Mathematics 13*(2), 221–252.

Chazal, F., B. T. Fasy, F. Lecci, A. Rinaldo, A. Singh, and L. Wasserman (2015, March). On the bootstrap for persistence diagrams and landscapes. *Modeling and Analysis of Information Systems 20*(6), 111–120.

Cohen-Steiner, D., H. Edelsbrunner, and J. Harer (2007). Stability of Persistence Diagrams. *Discrete Comput. Geom. 37*(1), 103–120.

Edelsbrunner, H., D. Letscher, and A. Zomorodian (2002). Topological Persistence and Simplification. *Discrete & Computational Geometry 28*, 511–533.

Ghrist, R. (2008). Barcodes: The Persistent Topology of Data. *Bulletin of the American Mathematical Society 45*(1), 61–75.

Halvorson, H. (2013). The Semantic View, If Plausible, Is Syntactic. *Philosophy of Science 80*(3), 475–478.

Hatcher, A. (2002). *Algebraic Topology*. Cambridge UP, Cambridge.

Manders, K. (2008). The Euclidean Diagram (1995). In *The Philosophy of Mathematical Practice*. Oxford: Oxford University Press.

Mumma, J. (2010). Proofs, Pictures, and Euclid. *Synthese 175*(2), 255–287.

Nicolau, M., A. J. Levine, and G. Carlsson (2011). Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival. *Proceedings of the National Academy of Sciences 108*(17), 7265–7270.

Perea, J. A. and J. Harer (2015). Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. *Foundations of Computational Mathematics 15*(3), 799–838.

Rosenstock, S. (2019). *A categorical consideration of physical formalisms*. Ph. D. thesis, UC Irvine.

Rosenstock, S. (2021). Clustering Schemes for Diverse Data Models. *Unpublished Manuscript*.

van de Weygaert, R., G. Vegter, H. Edelsbrunner, B. J. T. Jones, P. Pranav, C. Park, W. A. Hellwing, B. Eldering, N. Kruithof, E. G. P. p. Bos, J. Hidding, J. Feldbrugge, E. ten Have, M. van Engelen, M. Caroli, and M. Teillaud (2011). Alpha, Betti and the Megaparsec Universe: On the Topology of the Cosmic Web. In M. L. Gavrilova, C. J. K. Tan, and M. A. Mostafavi (Eds.), *Transactions on Computational Science XIV: Special Issue on Voronoi Diagrams and Delaunay Triangulation*, pp. 60–101. Berlin, Heidelberg: Springer Berlin Heidelberg.

Weatherall, J. O. (2017). Categories and the Foundations of Classical Field Theories. In E. Landry (Ed.), *Categories for the Working Philosopher*, pp. 329–348. Oxford University Press.

Perspectives on Causal Specificity

Abstract

Causal specificity is a measure of how important a cause is relative to another. Waters (2007) has developed a theory of causation that deals with specificity. Weber (2006, 2017a, 2017b) has thoroughly criticized it. I defend Waters's theory by showing that non-systematicity is unproblematic. I also argue that Weber's desiderata for theories of causation are too restrictive and insensitive to developments in biological technology. I finally challenge the most fundamental assumption in the framework of causal specificity—that bijective functions are most specific—thus calling for its reassessment.

Word count: 4,768

Introduction

Philosophers of biology debate whether some causes are more important than others in an explanation. The proponents of causal parity maintain that all causes of an effect are equally important. By contrast, the advocates of causal privilege argue that some causes are more important than others. According to the latter, a number of causes may be privileged and the degree to which each is privileged relative to others can differ (Waters, 2007; Weber, 2006; Woodward, 2003). I will explain the important aspects of this discussion using Waters's (2007) terminology.

Waters specifies conditions that a cause must meet in order to 'make a difference' in an effect (e.g., protein synthesis). The basic idea is that if a cause accounts for the variation in an effect (which occurs in a real population), then it is a difference maker with respect to that effect. Causes that counterfactually could have made a difference but that do not actually do so are called potential difference makers. Causes that *do* make a difference are called *actual* difference makers. If only one cause makes the difference, it is called *the* actual difference maker. If a number of causes make the difference, then each is *an* actual difference maker. Causes that actually make a difference are clearly more important than those that do so only potentially, since they account for the actual variation in the effect. The particularly difficult issue concerns the *degree* to which some given *an* actual difference maker is privileged relative to another. More precisely, let A:{x, y, z} denote the set of causes that actually make a difference in effect B. The crucial question is whether, say, x is privileged over y and z by degrees p and q, respectively. If so, we need to understand what this 'degree' is.

The degree of causal privilege is measured in terms of *causal specificity* (defined in section I), a concept developed by Weber (2006) and Woodward (2010). The central idea is that the more 'closely' the values of the cause variable map to those of the effect variable, the more specific the

said cause is with respect to the effect (as compared with its other causes). Waters (2007) claims that the causal specificity of the DNA is greater than that of the splicing agents vis-à-vis their common effect, mature mRNA. As a result, the DNA is the most privileged cause in this context. Although he largely supports the general framework of causal specificity, Marcel Weber (2006, 2017a, 2017b) has systematically criticized Waters's particular theory of causation. Specifically, he has argued that the theory's focus on actual populations often prevents it from being systematic.

I aim to accomplish three tasks in this paper. The first is to defend Waters's theory of causation against Weber's criticisms. The second is to examine the desiderata of theories of causation and argue that Weber's conditions seem unreasonable. The third is to criticize a fundamental assumption about causal specificity that its proponents share. In the first section, I outline Weber's (and Woodward's) account of causal specificity and provide some empirical details from molecular biology to contextualize the discussion. In section two, I explain Weber's criticism that Waters's theory is not systematic. I defend Waters by showing that this problem is not unique to his theory but is a feature of the explananda that all theories of biological causation need to contend with. [1,2] In the third section, I discuss some general issues around the desiderata for theories of causation and argue that Weber's conditions are unreasonable. The fourth section is concerned with criticizing the widespread assumption that bijective functions are causally most specific; I demonstrate that a type of non-bijective function can be more causally specific, thus

[1] I am focusing on Waters's (2007) theory in part because it presupposes Woodward's (2003), so a successful defense of the former entails a defense of the latter (cf. Weber, 2006).
[2] Despite Weber's critique over the last decade or so, neither Waters nor Woodward has responded to him except for Woodward's quick mention of Weber (2006) in a footnote (2010, 305, footnote 17). At present, none of Weber's papers on 'causal specificity' is cited by either Waters or Woodward.

suggesting that the framework's core assumption be reassessed. In the final section, I briefly highlight some implications of my arguments for debates in philosophy of biology.

I. Causal specificity and its relation to molecular biology

Causal specificity is defined in terms of 'functional mapping.' Suppose cause C and effect E are discrete variables, each ranging over finite sets of values. The causal specificity of C corresponds with how 'closely' its values map to those of E. More precisely, suppose that C and E range over $\{c_1, c_2, \ldots, c_m\}$ and $\{e_1, e_2, \ldots, e_n\}$, respectively, and the function f maps C-values to E-values (Weber, 2006; cf. Woodward, 2010).[3] The larger the number of C-values that map to their corresponding E-values, the more specific C is as a cause of E (as compared to its other causes). If each and every C-value maps to one and only one E-value, then the mapping is bijective, making C the most specific cause of E (Woodward, 2010).[4]

The debate about the relative specificities of various causes emerged in the context of molecular biology. In eukaryotes, sequences of nucleotides on the DNA are transcribed into the pre-mRNA using certain enzymes, such as the RNA polymerase. The pre-mRNA molecule is broken down and reconstituted using 'splicing agents,' which work in conjunction with other enzymes and background 'cellular machinery.' The parts of the pre-mRNA that are excised are known as introns, while its remaining parts are known as exons. The exons are combined to constitute what is then called the mature mRNA, a molecule used in protein synthesis.

[3] The function presupposes Woodward's (2003) manipulability theory of causation, according to which it should answer counterfactual questions such as, what would happen if C-value changes from c_1 to c_{25}. If f is counterfactually robust—as causal generalizations in biology should reasonably be—then E-value would change to e_{25}.

[4] I am assuming that the sizes of the two sets are identical (i.e., $m = n$).

Waters (2007) claims that the DNA and splicing agents both make a difference in the mature mRNA molecule, but the former is causally *more* specific. That is, the number of DNA-values mapping to the mature mRNA values is greater than the number of splicing agents' values mapping to those of the mature mRNA. Therefore, genes are the most specific or privileged cause of the mature mRNA.

II. Systematicity of Waters's theory of causation

Weber (2017b) argues that Waters's theory is not systematic, because it focuses on *actual* populations. He points out that the frequency of a given causal variable differs radically from one biological context to the next. For example, the frequency of splicing agents is much higher in eukaryotes than in prokaryotes. As a result, their causal specificity is significantly greater in the former than in the latter. Waters's theory purports to answer *general* questions, such as whether the causal specificity of the DNA is higher than that of the splicing agents. However, its focus on actual populations precludes it from treating a causal variable in a systematic fashion, thereby preventing it from answering general questions. Weber writes, "The main problem is that [Waters's theory] is very sensitive to the relative abundance of a causal factor in some defined population. Thus, [the causal variable's] values will be highly context dependent, to such an extent as to make any kind of systematic comparison across contexts difficult" (2017b, 578). The theory cannot therefore be used to answer general questions about a variable's causal specificity, and consequently, it does not allow systematic, cross-contextual comparison between the specificities of two (or more) causal variables. No doubt, variables whose frequencies *are* relatively uniform across biological contexts (e.g., RNA) do not pose problems for Waters's theory. So, to be precise, Weber's claim is *not* that Waters's theory can never treat a variable systematically. Instead, his

argument is that it cannot systematically analyse a causal variable whose frequency radically differs between contexts.

According to the advocates of causal privilege, the central purpose of developing theories of causation is to explain why biologists choose certain causes over others when explaining some phenomena that occurs in *real* biological populations (Waters, 2007; Woodward, 2003). Now, it is *not* a feature of Waters's theory that the frequency of splicing agents differs between eukaryotes and prokaryotes. Rather, this is a biological fact that all theories of causation must contend with. Consequently, there is nothing distinctive about Waters's theory that warrants the charge of non-systematicity. Consider a simple analogy. All empirical scientific theories face the problem of induction, the drawing of universal generalizations on the basis of finite evidence. This problem makes scientific theories fallible in principle, because there could exist some evidence that refutes the theory (Nola & Sankey, 2014). However, it would be peculiar to exclude some of the theories on this basis but spare others.

If Weber's criticism is to hold water, he needs to show that either (i) Waters's account fails to be systematic as a result of its own theoretical shortcomings, or (ii) there is at least one other theory that systematically analyses causal specificity. If he attempts to show (i), he must do so by referring solely to Waters's theory, not to its explananda. As I have argued, he relies on the intractability of the explananda to criticize the theory. Alternatively, Weber may demonstrate (i) by showing (ii), because the latter entails the former: if a theory can systematically analyze causal specificity, then the inability of Waters's theory to do the same must be the result of its internal features (or of its application). In other words, if another theory can provide a systematic analysis, then this gives good reason for thinking that the explananda are tractable after all. Consequently, the failure of Waters's theory to do the same cannot be the result of its explananda. Its failure

would arguably be the result of its theoretical apparatus. However, Weber's criticism satisfies neither (i) nor (ii). Therefore, it does not provide reasonable grounds for deeming Waters's theory uniquely problematic.

III. Biological normality and desiderata for theories of causation

My purpose here is to extend the foregoing discussion by analyzing the conditions Weber thinks theories of biological causation must meet. I will argue that Waters's theory meets these conditions. I will also suggest that the conditions themselves are quite misplaced.

Weber (2017b) claims that any plausible theory of causation must meet the conditions of 'biological normality.' A causal intervention is biologically normal if it (a) results from natural processes with non-negligible probabilities and (b) is compatible with the ordinary biological functions of the organism. For instance, DNA transcription is a natural process with non-negligible probability and is compatible with the organism's functions. I think Waters's (2007) theory meets these conditions. The causal interventions in actual biological populations are by definition natural and compatible with the organism's functioning. It is impossible to consider normal biological populations without also thinking about their causal relations as natural and compatible. Alternatively, the concept of actual populations would be vacuous, if not self-contradictory. Accordingly, Waters's theory cannot fail to meet these conditions, because it explicitly focuses on and restricts itself to actual populations. Hence, the theory satisfies the conditions its critic thinks any acceptable theory of biological causation should meet.

A particular reading of Waters (2007) makes it even more difficult for his theory to fail to meet Weber's conditions. The *conditional* reading begins with the fact that biologists consider certain causes as more important than others when explaining some phenomena. The task of

Waters's theory, on this reading, is to provide a principled account that explains their choices. In particular, its task is to answer questions of the following kind: given that a number of causes account for the variation in the effect, which of these is causally most specific? All of the ordinary causes a biologist invokes when explaining some phenomena are natural causes compatible with the organism's functioning, thus satisfying Weber's conditions (a) and (b), respectively. The conditional reading takes this as a given, and the purpose of the theory, on this reading, is to provide a principled account of the relative importance of various causes. Hence, on the conditional reading, Waters's theory will always meet Weber's conditions.

Let us turn to a more general discussion of Weber's conditions and examine whether these demands capture the intuitions about theories of causation that philosophers of biology have in mind. I will focus on Weber's first condition, which states that a causal intervention is biologically normal if it (a) results from natural processes with non-negligible probabilities. This requirement is problematic for at least three reasons. First, natural processes with *negligible* probabilities remain philosophically unanalyzable even though they are ordinarily thought of as biologically normal. For instance, (successful) genetic mutations have very low probabilities. However, they are causally significant for explaining a wide variety of biological phenomena, such as phenotypic variation. In Waters's terminology, genetic mutations are causes that make a difference, and biologists no doubt invoke them in their explanations. Yet, if condition (a) is accepted, genetic mutations (and other natural processes with negligible probabilities) would remain unanalyzable from the perspective of a philosophical theory of causation. In short, this condition wrongly excludes improbable factors that are nonetheless causally relevant.

The second reason that this condition is problematic is that theories of causation which satisfy it will not analyze non-natural interventions. Interventions are usually thought of as "non-

natural" if they cannot occur without the use of technology. Nonetheless, some non-natural interventions, such as *in vitro* fertilization or genomic editing, are often highly relevant for explaining biological phenomena. Indeed, the advent of technology allows biologists to intervene in very specific ways. A philosophical theory that fails to analyze this excludes important aspects of biological phenomena. Weber's condition has precisely this effect: it deems as unacceptable philosophical theories that analyze non-natural but causally relevant interventions. Once again, the condition is too restrictive and, importantly, it is insensitive to developments in biological technology.

Third, Weber's condition is problematic because the boundaries between 'natural' and 'artificial' are more difficult to define than he assumes. For example, gene editing is arguably non-natural because it is carried out using technology. However, edited genes are transcribed and translated into proteins using 'natural' processes. Only the initial cause in this chain of events is supposedly non-natural. The subsequent causes are perfectly natural. Indeed, gene editing technology meets the second condition that (b) the interventions be compatible with the rest of the organism's functioning. However, by requiring that interventions be natural, condition (a) precludes theories of causation from selecting causally relevant factors in technologically altered populations. In other words, on this condition, theories of causation will altogether ignore populations with 'non-natural' interventions even if these interventions are causally fundamental (e.g., gene editing). Consequently, the condition is, once again, unduly restrictive and insensitive to developments in biological technology.

Furthermore, Weber's conditions are in tension with his criticism (presented in the previous section) that Waters's theory fails to treat a causal variable systematically. On the one hand, his condition (a) requires theories of causation to focus only on *actual* populations. On the other hand,

his criticism of Waters's account suggest that he requires theories of causation to be systematic, implying that they should *not* be susceptible to the changing reality of actual populations. A more charitable interpretation is that he requires the theories to be systematic despite the mutability of biological phenomena. However, as I argued in the previous section, it is difficult to see how a theory can be systematic when its explananda is highly mutable. This is surely the case with biological populations, as Weber himself claims. As a result, it is difficult to see how theories can restrict themselves to actual populations *and* be systematic. It seems, then, that these two requirements are in tension with one another. Minimally, the desiderata for theories of causation must be mutually agreeable. They would otherwise require theories to perform incompatible tasks. Weber's requirements (of biological normality and systematicity) fail to meet even this demand.

In light of these considerations, it is natural to ask Weber to explain how these two requirements are compatible. There are at least two courses of action available to him. First, he may argue that while no actual theory of causation has succeeded in satisfying these requirements, it is possible that some theory *could* meet them. Second, he may develop a theory that satisfies both requirements. However, no actual theory of causation (that I am aware of) meets these requirements, suggesting that they are overly restrictive. This rules out the second course of action. As for the first course, it needs to be shown how exactly a theory could be systematic and focus on actual populations. It is insufficient to stipulate conditions without providing at least a sketch of a possible solution. Weber has not provided a sketch of any kind.

Finally, I want to return to the requirement of systematicity. I have already suggested that this requirement is overly demanding. Nevertheless, even if we assume for argument's sake that it could be satisfied, the new insights gained from a systematic analysis of causal specificity do not sufficiently advance our understanding of biological causation. Generally speaking, philosophers

of biology no longer maintain that biological generalizations are universal. They recognize that the generalizations only distribute over well-defined domains (Waters, 2007). In my view, the case of causal specificity should be understood within this framework; there is no need for cross-contextual, domain-general comparisons when domain-specific analyses suffice. For instance, it can be meaningfully asked whether the causal specificity of the DNA with respect to some effect is greater than that of the splicing agents with respect to the same effect within a *well-defined population*. If 'yes,' then we have reason for thinking that, in this particular context, the DNA is more causally important than the splicing agents with respect to a given effect.

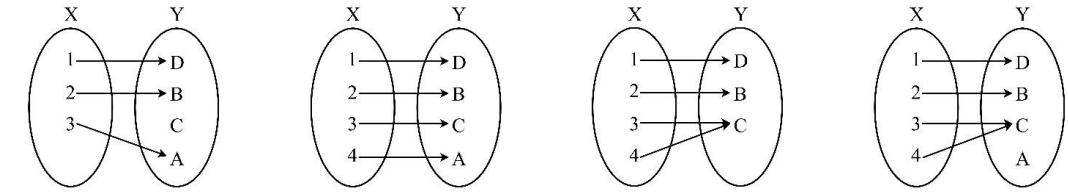
However, systematic analyses do not provide insights of this type. In particular, it is not very informative to compare the specificities of two (or more) causal variables with respect to different effects (in same or different contexts) or same effects (in different contexts). The former are uninformative because there is no relevant commonality based on which the differences could be meaningfully compared; the latter only tell us that a certain cause is more specific than its counterpart with respect to the same effect in a number of contexts. Yet, as Weber rightly points out, the frequencies of variables radically differ between biological contexts. So even if a systematic analysis generated new insights, it would remain largely uninformative when applied to genuine populations. In light this, it is clear that domain-general or systematic analysis is usually uninformative. Yet, that is precisely what the requirement of systematicity demands from theories of causation. Because this requirement is not conducive to advancing our understanding of causation in biology, it is best to altogether eliminate it.

IV. Are bijective functions (causally) most specific?

The aim of this section is to challenge the fundamental idea in the framework of causal specificity: that bijective functions are causally most specific. I will introduce some technical terminology before explaining that a type of non-bijective function can be more specific.

In the first section, I explained specificity in terms of mapping between variables. Let *f* be a function that maps *C*-values to *E*-values. *f* is considered a function if and only it meets the following conditions. First, all *C*-values must map to some *E*-value; there cannot be unmapped *C*-values. Second, no *C*-value can map to more than one *E*-value; each *C*-value must map to at most one *E*-value. Nothing about *E*-values is relevant when determining whether *f* is a function.

A *surjective* ('onto') function is one in which all *E*-values are the image of some *C*-value under *f*, meaning that there are no unmapped *E*-values. A function is *injective* ('one-to-one') when each *E*-value is the image of at most one *C*-value under *f*. That is, given that some *E*-values are mapped, each of these is mapped by at most one *C*-value. A function is *bijective* if and only if it is surjective and injective. The following diagram summarizes these concepts:



An injective non-surjective function (injection, not a bijection) | An injective surjective function (bijection) | A non-injective surjective function (surjection, not a bijection) | A non-injective non-surjective function (also not a bijection)

This framework can be used to map genuine causal relations in biology and to determine their relative specificities. The table below presents some *bona fide* causal relations alongside their respective mappings (Weber, 2017a, 17-8):

Stage of gene expression (C→E)	Mapping *f*(C)=E
1: DNA → DNA (replication)	Bijective
2: DNA → RNA (transcription in prokaryotes)	Bijective
3: RNA → DNA (reverse transcription)	Bijective
4: DNA → pre-mRNA (transcription in eukaryotes)	Bijective
5: pre-mRNA → mature mRNA (in eukaryotes)	Not a function[5]
6: exon parts → protein domains	Surjective non-injective
7: mature mRNA → proteins (translation in eukaryotes)	Surjective non-injective

In this framework, Weber favors a numerical interpretation of the specificity of these mappings. He writes, "Depending on the range of invariance and the *number of values* that the independent [*C*-values] and dependent [*E*-values] variables can take, we can speak of a relation being more or less causally specific" (2006, 606; my emphasis). He claims that "[t]he elements in the codomain [*E*-values] may be mapped onto by different *number* of arguments [values] from the domain [*C*-values] (in the surjective and non-injective cases), or different *proportions* of elements in the codomain may be mapped onto by an argument from the domain (in the injective and non-surjective cases)" (2017a, 16; my emphasis). But in the same paper he distances himself from a proportional notion when he writes, "By "causally most specific" I mean that genes bear Woodward's relation INF [influence] to proteins in the highest degree. By "the highest degree" I mean that the *number of values* that the variables on both sides of the INF relation can take is vastly higher (i.e., many orders of magnitude) than that of any other causal variables that bear the

[5] Because parts of the pre-mRNA are excised (introns), some of its values cannot map to those of the mature mRNA. Consequently, pre-mRNA to mature mRNA mapping does not constitute a function.

relation INF to protein sequences (e.g., splicing agents)" (2017a, 32; my emphasis). Weber clearly favors a numerical conception of causal specificity, according to which one ought to count the number of mappings between *C*-values and *E*-values. The greater the number of mappings between *C* and *E*, the more specific the former is as a cause of the latter (as compared with its other causes). The functions that arguably exhibit the highest number of mappings are bijective, leading to the consensus view that they are causally most specific (Weber 2017b; Woodward, 2010).

I want to argue that surjective non-injective functions can be more specific than bijective functions. If this is true, then the assumption that the DNA is causally most specific may need to be reevaluated. To be sure, I am not claiming that the DNA *actually* fails to be the most specific cause. Instead, my argument will try to show that the consensus view—that bijective functions are always causally most specific—is not correct.

Consider a cause F:{1, 2} and its effect Z:{a, b}. Suppose the function m:$F \rightarrow Z$ bijectively maps F to Z. Consider another cause G:{1, 2, 3} and its effect Z:{a, b}. Suppose the function n:$G \rightarrow Z$ surjectively non-injectively maps G to Z. Let $G(x)=F(x)$ but let $G(3)=G(1)=F(1)$, meaning that *G*-values '1' and '3' and *F*-value '1' map to *Z*-value 'a.' Calculating causal specificities using Weber's approach generates values 2 and 3 for *m* and *n*, respectively. That is, two values of *F* and three of *G* map to Z, meaning that the latter (surjective non-injective) is more specific than the former (bijective). Thus, the same effect can have two causes such that the one which maps to it bijectively—which is presumably the most specific mapping—is *less* specific than the one that maps to it surjectively non-injectively. This is particularly problematic for the proponents of causal privilege, because they regard some of the higher-order causal relations in biology (e.g., mature mRNA → proteins) as involving surjective non-injective mappings. They also consider the most fundamental and specific causes (e.g., DNA → primary transcript) as bijective.

Bijective functions being less causally specific than surjective non-injective functions provides a reason for re-examining the framework. The important intuition behind the idea that bijective functions are supposedly most specific is that nothing seems more specific than one cause giving rise to one and only one effect. This intuition seems *prima facie* correct. Nonetheless, the framework allows non-bijective functions to be more specific. Consequently, the issue is likely with the "number of values" conception used here. As a result, alterations may be required to make this approach work. One suggestion is that the mappings could be used without requiring that they be functions. This would provide more room for theoretical development. More radically, an altogether different approach may be developed that (quantitatively) captures the idea of causal specificity (cf. Griffiths et al., 2015).

V. General implications and conclusion

I will briefly highlight some implications of my arguments. First, the proponents of causal privilege (Waters, 2007; Weber, 2006; Woodward, 2003, 2010) agree that some causes are more important than others when explaining some phenomena. While there are differences about which cause matters to what degree in some context, the authors agree that some causes are definitely privileged over others. Waters's (2007) theory, which relies on Woodward's (2003) manipulability theory of causation, is one of the most thoroughgoing attempts at developing a theory of causal selection in biology. Its central purpose is to provide an alternative to causal parity, according to which all causes in an explanation are equally important. In this paper, I explained the criticisms of Weber, a proponent of causal privilege, against Waters's theory. If my defense of the latter is successful, however, it may go some way in resolving issues internal to the framework of causal privilege. More optimistically, this defense could unify the various proponents of causal privilege, and a robust account of causal privilege as an alternative to causal parity may be developed.

Second, I argued that Weber's conditions for theories of biological causation are overly restrictive. I suggested that these conditions should not prevent theories of causation from analysing causal interventions made using technology. Given the widespread increase in genetic technology, it seems misplaced to regard as irrelevant factors that are causally significant in biological explanations. As such, the desiderata for theories of biological causation may need to be reassessed.

To conclude, I explained what causal specificity is and how it relates to molecular biology. I tried to rebut Weber's argument that Waters's theory fails to be systematic. I also showed that the theory meets Weber's conditions of biological normality, and I argued that his conditions are quite unreasonable. Finally, I showed that surjective non-injective functions can be more specific than bijective functions, thus suggesting the need for re-evaluating at least this tenet in the framework of causal specificity.

References

Griffiths, P. E., Pocheville, A., Calcott, B., Stotz, K., Kim, H., & Knight, R. (2015). "Measuring Causal Specificity." *Philosophy of Science*, 82 (4), 529–555.

Nola, R., & Sankey, H. (2014). *Theories of Scientific Method: An Introduction*. Routledge.

Waters, C. K. (2007). "Causes that make a difference." *The Journal of Philosophy* 104 (11): 551-79.

Weber, M. (2006). "The Central Dogma as a Thesis of Causal Specificity." *History and Philosophy of the Life Sciences* 28 (4), 595–609.

---. (2017a). "Causal Selection versus Causal Parity in Biology: Relevant Counterfactuals and Biologically Normal Interventions." Forthcoming in: C. Kenneth Waters and James Woodward (eds.), *Philosophical Perspectives on Causal Reasoning in Biology*.

---. (2017b). “Which Kind of Causal Specificity Matters Biologically?” *Philosophy of Science*, 84 (3), 574–585.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

---. (2010). “Causation in biology: Stability, Specificity, and the Choice of Levels of Explanation.” *Biology & Philosophy*, 25 (3), 287–318.

Creativity in the social epistemology of science

Mike D. Schneider*†

Abstract

Currie (2019) has introduced a novel account of creativity within the social epistemology of science. The account is intended to capture how conservatism can be detrimental to the health of inquiry within certain scientific communities, given the aims of research there. I argue that recent remarks by Rovelli (2018) put pressure on the applicability of the account. Altogether, it seems we do not yet well understand the relationship between creativity, conservatism, and the health of inquiry in science.

*To contact the author, write to: Mike D. Schneider, Center for Philosophy of Science, University of Pittsburgh; e-mail: schneider.michael.d@gmail.com.

†I would like to thank Kino Zhao and Kyle Stanford for their helpful conversations and comments in the planning of this paper, as well as Jim Weatherall, Kyle Stanford (again), and Adrian Currie for their input in my revising it.

1 Introduction

Currie (2019) argues that research in existential risk ('X-risk') should be more creative than it likely is, given the realities of contemporary scientific practice. In the course of the argument, he introduces a general account of creativity in scientific discovery (hereafter, 'creativity'). This account is intended to capture how conservatism can be detrimental to the health of inquiry in scientific communities, given certain aims of research. It is also advertised as complementing the use of formal modeling in studying policy initiatives within the social epistemology of science.

Independent of Currie's project, Rovelli (2018) decries a "why not?" ideology he reports is in vogue within his scientific community, engaged in fundamental physics research. By his reckoning, this ideology promotes a method of guesswork. His concern is that such a method is detrimental, given facts about his community and their research aims.

Here, I will argue that Rovelli's remarks, when interpreted in light of Currie's account, raise trouble for the general applicability of the latter. Evidently, Currie's account fails to countenance the possibility that revolutionary theorizing might be valuable, as features in Rovelli's argument. But since it is difficult to discern when revolutionary theorizing is likely not valuable to a community, it is unclear when Currie's account may be deemed appropriate for studying the effects of conservatism on the health of inquiry therein. This threatens to undermine the use of such an account in arguments undergirding policies meant to respond to conservatism. It would be prudent to seek out means of identifying what it is about any given scientific community that could render Currie's account appropriate there.

2 Creativity in science

Stanford (2019) has argued that the structures and institutions of contemporary science foster conservatism in research, stifling revolutionary theorizing. Currie (2019) is concerned that the same conservatism is detrimental to inquiry within X-risk. This is because, according to Currie, disciplines like X-risk are best pursued creatively. Arguing that creativity is in tension with conservatism, Currie concludes that the scientific communities focused on disciplines like X-risk are likely insufficiently creative— the structures and institutions of science stack the deck against the disciplines' prospects.

As just presented, Currie's project depends essentially on his providing an explicit account of creativity within a scientific community. The remainder of this section is dedicated to describing the account he provides, as well as developing it further (where necessary) in a friendly manner.

Consider the situation wherein there is some well-posed problem, whose solution a scientific community agrees constitutes the aim of their collective research. The statement of the problem places severe constraints on what counts as viable research within that community, united by that aim. We may think of the statement of the problem as characterizing the research program pursued by that community. And associated with that problem is, following Currie, a collection of possible solutions. This 'solution space' is meant to be roughly coextensive with all professional moves available to members of that community, engaged in that research program. The researchers occupy points in the solution space, and they choose which points to occupy next.[1]

[1]In fact, there are other professional strategies that are ultimately available to researchers, regarded as decision-making agents. Whether activity gets channeled into those other strategies, rather than into moving between solutions, is an important degree of freedom in Currie's account.

Currie introduces into this picture the following two metaphors. 'Hot searches' through solution space are energetic; 'cold searches' are the opposite. A hot search refers to a sequence of points, whose iterative selection by a theorist describes that theorist as hopping around through the solution space. A cold search refers to a similar sequence of points, except that it describes the behavior of a theorist who is nearly staying still.

To make these metaphors, Currie needs a notion of distance between points in the space. He borrows from Bayesian epistemology to develop one. (I will have more to say that is critical of this below.) By his reckoning, distances to solutions are relativized to each individual at a time, and are indexed to that individual's credences at that time. So, roughly speaking, solutions assigned low priors are far, and solutions assigned high priors are near.[2]

Currie does not elaborate on the interpretation of these priors. Evidently, he has in mind something pragmatic: "Our priors serve to set expectations across a space of possible solutions to a problem" [p. 6]. In this respect, the account is non-committal about what it is that ultimately makes a solution worth visiting. We are free to suppose that there is some unspecified constellation of virtues, possibly specific to the research program at hand, that one hopes is jointly maximized (i.e. via some method of aggregation) by whatever solution is visited next. On this picture, hot searches are sequences for which the researcher's decisions are insensitive to their beliefs about where it will be prudent to visit. Oppositely, cold searches occur when the researcher's choices correlate strongly with those beliefs.

Currie then defines an agent's creativity in terms of their propensity for hot searches.

[2]As will become clear, it may be that we ought to insert a ceteris paribus clause here. If so, we would say that whatever are otherwise the distances to solutions, those numerical values are then systematically deformed to reflect comparative facts about one's priors over each solution.

In other words, an agent is creative in proportion to the unconditional probability that they attempt a distant, low-credence solution. A community's creativity, meanwhile, is defined to correspond with what would generally occur if the members of the community were all individually creative. The upshot is that a community's creativity is defined as proportional to the efficiency with which they explore solution space widely. (What it means to explore widely is, of course, agent-relative. Here, we might assume that a community explores widely when it does so by the lights of most of its members.)

This wide exploration of solution space is in contrast with what, following Currie, we may call 'pooling'. Intuitively, pooling occurs when individuals within the community fail to be creative, each favoring cold searches instead of hot searches. But, as Currie notes, pooling may be avoided in such a case, provided that the community is cognitively diverse. So long as cognitive diversity is understood in terms of diverse distributions of priors, cognitively diverse individuals engaging in cold searches will, collectively, explore widely. This community would count as creative, according to Currie, even though the individuals who comprise it do not.

The creativity of a community is therefore not uniquely determined by facts about the creativity of its constituents. Their propensity for peer disagreement (and so, the social structure of science, etc.) also matters. And according to this view, a community may be made more creative in various ways. One way is by interventions to promote sustained cognitive diversity, as we have understood it here. Another is by incentivizing hot searches, or increasing creativity at the individual level. In both cases, pooling is reduced, in favor of wider exploration.[3]

[3]A third way to increase creativity, noted by Currie, is to impose on the community a diverse collection of search algorithms. But this raises a question: what distinguishes, in practice, our imposing a diverse collection of search algorithms from our incentivizing hot searches? At the

Building on recent work by Stanford (2019), such interventions are, according to Currie, in contrast with the unchecked effects of conservatism in professional science today. This is because, according to Currie, conservatism promotes pooling, as we have understood it here. But depending on the given research program, it may or may not be detrimental that science today is, generally, conservative. This is because a research program ought to be assessed individually, according to the "local details" [p. 3] relevant to it. Those details determine, for instance, whether the community is better off investing in strategies other than those relevant to scientific discovery (cf. footnote 1 above). If so, any resulting pooling according to shared priors need not be unhealthy.

As just stated, the utility of Currie's account is ultimately going to rest on certain further facts: which kinds of local details ought we to recognize as rendering creativity— as opposed to pooling— a standard of good epistemic health in the community? Such local details are encoded, we may suppose, in the statement of the problem that constitutes the aim of that community's research. Recall that it is from this problem that, in principle, we may extract the parameters of the solution space we envision the community to explore. It follows that assessments of the local details of a research program will generally shape our expectations about the solution space associated with the problem. Likewise, facts about a solution space can correlate with facts about whether pooling or creativity is preferred in the corresponding research program.

Unfortunately, Currie does not state how such a correlation would work. This omission could suggest that we ought not to regard local details as shaping our

level of analysis presently provided, it is unclear that there is any distinction. As suggested in footnote 2, it may be that we should ultimately think of solution space as admitting some intrinsic structure, independent of credences. In that case, search algorithms could be defined with respect to that intrinsic structure, and would generally result in searches that appear hot.

expectations about solution space (besides via shaping our priors). But this would render Currie's account in tension with the standard interpretation of formal landscape models. Currie regards the use of such models within the social epistemology of science as complementing his approach (cf. p. 11 in the article). In such models, one typically regards the intrinsic structure of the landscape as an independent variable, whose possible values encode arbitrary research environments. So too, we might conclude, the structure of a solution space should reflect facts about the corresponding research program.

In light of this, I think it is appropriate to regard Currie's discussion of X-risk as illustrating the reasoning that would shape the relevant solution space. His ultimate conclusion is that X-risk should be creative because it should be "multi-disciplinary, pluralistic, and opportunistic" [p. 26]. We might speculate, on the basis of this, that the local details relevant to the problem of X-risk render the solution space as unusually vast.[4] In a vast solution space, cold searches could seem unfruitful, no matter how cognitively diverse we may plausibly imagine are the researchers. Consequently, creativity is generally preferred in such a case, consistent with Currie's reasoning about X-risk.

To recap: treating research programs as solution spaces, creativity is a matter of how the relevant communities explore those spaces, given priors. Conservatism encourages pooling according to shared priors, which is opposite creative exploration. But specific facts about the solution space at hand can determine, in a given community, which of

[4]There is room for disagreement here. For instance, Currie's discussion of X-risk places some emphasis on its normative aspect— i.e. threat mitigation— and its role in the public eye. It is not clear what these would have to do with the size of the solution space. This ambiguity motivates a revisionist attitude toward distances in the space. (See also footnotes 2 and 3 above.)

creativity or pooling is likely preferred. Those facts are ultimately grounded in the statement of the problem identified by that community as constituting their research program.

3 The situation in fundamental physics

Consider now the article by Rovelli (2018). Rovelli is a theoretical physicist focused on quantum gravity, the problem that characterizes fundamental physics research today.[5] Indeed, we may understand the problem of quantum gravity to be that which shapes the relevant solution space, against which creativity in fundamental physics is to be assessed. In what follows, I take Rovelli to have expertise regarding that solution space, as well as privileged access to it.

Rovelli's article is adversarial. Our attention is best directed to a passage that comes in the middle, immediately following his presentation of what he calls the "why not?" ideology. According to Rovelli, this uncritical ideology is responsible for the rise of a damaging method of guesswork in contemporary fundamental physics practice. According to the method, reason need not be (nor can be, fruitfully) given to merit the study of any new research proposal. The criticism of the method proceeds as follows [p. 7]:

> Arbitrary jumps in the unbounded space of possibilities have never been an effective way to do science. The reason is twofold: first, there are too many possibilities, and the probability of stumbling on a good one by pure

[5]This is, of course, a massive simplification. But so too is the problem characterizing X-risk in Currie's project. Whether the simplification is tolerable despite such objections depends on the particular context of its use.

> chance is negligible; but more importantly, nature always surprises us and we, the limited critters that we are, are far less creative and imaginative than we may think. When we consider ourselves to be "speculating widely", we are mostly playing out rearrangements of old tunes: true novelty that works is not something we can just find by guesswork.

As in Currie's article, we have here a spatial account of scientific discovery. Scientists decide how to move amongst points in the space (now, of 'possibilities', rather than 'solutions'). The role of the "why not?" ideology is to support a method of guesswork. We can understand this method as a decision procedure, the repeated execution of which amounts to "arbitrary jumps" in the space. (More formally, we might think of such a method as analogous to Monte Carlo sampling, with respect to some unspecified probability distribution on the space. Based on the context surrounding the quoted passage, Rovelli clearly has in mind a distribution that is meant to be uncorrelated with one's priors.) But absent any greater detail about the account Rovelli envisages, it is unclear why such a method should be as damaging as he claims. Prima facie, Currie's account of creativity should be helpful as a means to interpret the argument.

In Currie's framework, Rovelli's 'space of possibilities' may be understood as a solution space for the problem of quantum gravity. The solutions to the problem are, then, candidates for what may turn out to be a satisfying theory of quantum gravity. Given this reading, Rovelli's principal claim about the space is that it is vast. This seems right. In other contexts, this space is taken to be synonymous with 'theory space', the collection of all possible fundamental theories (see, e.g. (Dardashti, 2019)). From here onward, I will adopt this 'theory space' language when talking about the space of

solutions relevant to the problem of quantum gravity.[6]

Recall that creativity at the community level is spelled out, on Currie's account, in terms of exploring widely in the relevant solution space. I have suggested that we understand Rovelli's remarks in terms of fundamental physicists exploring the vast theory space corresponding to the problem of quantum gravity. Since the space is vast, by the argument at the end of the previous section, creativity is likely preferred to pooling. In other words, a more creative community is likely better off, given the local details of the problem of quantum gravity. Wider exploration should be good here.

Meanwhile, fundamental physicists are, according to Rovelli, *uncreative* (or, at least, are "far less creative" than they may think).[7] On the present interpretation, this would suggest that fundamental physicists fail to explore widely. Increasing creativity should be desirable.

Naively, guesswork is one such method to do so. (As described above, except if the sampling is with respect to a probability distribution correlated with one's priors, guesswork will generally produce hot searches.) On Currie's account, we may thereby understand Rovelli to hold the view that the method of guesswork happens to be implemented poorly by his community. Moreover, according to Rovelli, when his community engages in guesswork, they fail to speculate as "widely" as they typically believe themselves to speculate. So: the community does not explore widely, and they fail to recognize that this is the case.

This seems to provide a sufficient reason that the method is, according to Rovelli,

[6]In (Schneider, 2020), I criticize the relevance of this 'theory space' view in assessing the methodology of quantum gravity research.

[7]What relation this testimony could bear to the broader conversation about conservatism in science is interesting to consider, but a tangent at present.

damaging. Because theory space is vast, creativity constitutes a standard of good epistemic health in contemporary fundamental physics. Meanwhile, the community's poor implementation of guesswork fosters an exaggerated perspective as to how healthy their inquiry really is. Our initial hunch was correct: Currie's account of creativity can help us get traction on Rovelli's argument.

Yet, there is something unsatisfying about this interpretation of the argument. Consider the reason that Rovelli supplies for his testimony that the community implements the method of guesswork poorly. The poor implementation is due to the fact that "we, limited critters that we are, are far less creative and imaginative than we may think". In other words, guesswork is implemented poorly by his community, because their being limited ensures that they *cannot implement it well*. In particular, it his community's lacking creativity (and imagination), on this interpretation, that ultimately bears responsibility for the method being damaging.

Whether Rovelli's argument is compelling, so interpreted, is therefore going to turn on whether a community's lacking creativity can be understood to intervene on the efficacy of a method they attempt to employ. And here, Currie's account provides little guidance. Facts about the community's pooling with respect to shared priors cannot obviously prohibit researchers, all of whom are willing to speculate irrespective of their priors, from doing so. In this respect, Rovelli's argument depends on creativity (or the lack thereof) playing a further role in the social epistemology of his community than is readily countenanced by Currie's account.

Note that this observation does not present an objection to Currie's argument, as his argument does not require that his account of creativity be complete. Nonetheless, as I will now discuss, Rovelli's argument is ultimately compelling, provided that we attribute

to Rovelli the view that revolutionary theorizing is valuable in contemporary fundamental physics. And recognizing the importance of such a view to Rovelli's argument should make us wary about assertions that Currie's account is applicable in any particular epistemic situation. Currie's account cannot merely be assumed to capture how to assess the epistemic impact of conservatism on a research program, for which creativity is healthy. A further question about whether or not revolutionary theorizing is valuable complicates the assessment.

4 Revolutionary theorizing and the health of inquiry

Suppose that there exist possibilities in theory space that are assigned prior probabilities of zero by all members of the community. Whereas many possibilities are *accessible* to the community, in virtue of being assigned non-zero priors by someone, these further possibilities are *inaccessible*. On Currie's terms, these are possibilities that are located an infinite distance away from the community, and are regarded as infinitely less promising to visit than any accessible possibility.[8]

In such a case, no matter how creative the community is regarding the accessible possibilities, some of theory space will never be explored. So, provided that guesswork fails to be defined over inaccessible possibilities, the method could fail to spread the community as wide as might, ultimately, be desired. This idealized setup sounds

[8]Assignments of zero-probability priors to non-contradictions are antithetical to an orthodox Bayesian epistemology. So, it is not obvious that the present supposition, in the case of theory space, is faithful to Currie's project. Nonetheless, given some other structure to the space (cf. footnotes 2-4), we may understand zero-probability priors as an idealization that "pushes off to infinity" the corresponding possibilities. They are, in effect, disconnected from the accessible ones. No amount of information gleaned from work on the latter could ever reign them in.

promising as a means to recover why, according to Rovelli, his community cannot implement guesswork well. We need only to attribute to Rovelli two further claims. The first is that his community's lack of creativity results in there being some possibilities that are inaccessible. The second is that at least some of those inaccessible possibilities are important to the aims of his community's research.

Evidence that Rovelli would endorse each of these claims may be found within the passage already quoted. Namely, what is inadequate about guesswork, says Rovelli, is that it does not yield "true novelty that works". This is because employing it results (instead) in "playing out rearrangements of old tunes". If we interpret the rearrangements of old tunes as the accessible possibilities, his claim is this: what there is to be sought in fundamental physics— i.e. true novelty that works— in fact resides in the inaccessible part of theory space.

Suppose that this reading is correct, and what there is to be sought in fundamental physics is, according to Rovelli, presently inaccessible. Then it is a symptom of the community's not being creative, according to Rovelli, that the implementation of guesswork necessarily fails to engender wide enough exploration. This is because the relevant sampling procedures fail to be defined over the whole of what is *worth exploring*.

We have thereby found a means to articulate the lingering part of Rovelli's argument, which we were unable to do in the previous section. Namely, says Rovelli: what is worth exploring fails to be coextensive with the accessible part of theory space. As a result, guesswork is ineffectual. Worse, employing the method misleads the community in their self-assessment of whether they are sufficiently creative, consonant with their research aims. This is because the method only promotes wide exploration of a kind that is unsuitable for assessing the health of inquiry in fundamental physics. It only

countenances that which is *conceived as* worth exploring (i.e. rather than what *is*).

If this is how we are to understand Rovelli's argument, it is easy to generalize the lesson. Consider any context wherein one has reason to regard the accessible part of solution space as failing to include some of what is worth exploring (putting off, at least for another few paragraphs, the issue of what it means for something to be worth exploring). This is a context in which genuinely revolutionary theorizing is needed, which renders accessible more of the space. In other words, if a community has reason to value revolutionary theorizing in their research, no amount of hot searching amidst that which is conceivable will amount to healthy inquiry. This is despite creativity remaining a standard of good health in that community, given their research aims.

But such a conclusion spells trouble for the applicability of Currie's account in arguments about policy. Currie's observation, as discussed above, is that conservatism promotes pooling with respect to shared priors. To the extent that creativity is anticorrelated with such pooling, Currie concludes research programs that ought to be creative likely suffer, in virtue of conservatism. Therefore, interventions that would promote creativity in the relevant communities would be well motivated, given the broader context of science today. (Indeed, this is just what Currie calls for in the case of X-risk.)

But now, there is cause to doubt that creativity has anything to do systematically with pooling, as defined with respect to shared priors. Creativity may, for instance, be anticorrelated with an entirely different kind of failure to explore, measured against an entirely different distance measure on the space. At least when revolutionary theorizing is valued, this seems to be the case. Indeed, one might even imagine situations wherein pooling, as measured against priors, provides explicit means of playing with what it is

that we conceive as worth exploring. (Rovelli seems to have something like this in mind in his advocating for a method built on continuity, in order to break away from playing rearrangements of old tunes.)

If so, interventions to promote creativity cannot be motivated against a background of conservatism, at least as Currie has presented the topic. In cases such as these, we require a different sort of reason to motivate interventions in response to conservatism (when, still, creativity is important). For instance, suppose that the conclusion is warranted: conservatism deprives the relevant community of access to much of solution space (cf. footnote 7). Then it is plausible that what is sought by the community is inaccessible, in which case revolutionary theorizing might be valuable. Policies intended to promote creativity in that community could then be motivated, given the broader conservatism of science today. (And enacting such policies would be all the more important if, following Stanford, we further regard conservatism as stifling revolutionary theorizing.)

On the other hand, we might imagine some cases (perhaps that of X-risk) in which Currie's account adequately captures the effects of conservatism on inquiry. These are cases where we regard a community's capacity for revolutionary theorizing as, antecedently, unimportant to assessing the health of inquiry therein.

Such cases may arise in practice. But if they do, it is very difficult— if not impossible— to reliably identify them as such. What is up for grabs here is our epistemic access to whether that which we presently conceive as worth exploring happens to be coincident with that which is worth exploring. This is one lesson of Stanford's original project, which foremost concerned our means of evaluating the contemporary threats posed by the problem of unconceived alternatives. The upshot is that there may turn out

to be no problem inherent in the applicability of Currie's account in certain cases. Yet, there is a severe problem in asserting when we are reliably in such a case. This matters for the argumentative force of any call for new incentives to promote creativity in any particular community, based on his account. Namely, one must commit to the belief that, whatever it means for a solution to be worth exploring— i.e. given the ultimate aims of the community's research, the individuals' understandings of the problem that shapes that research, and so on— that solution is presently conceived as such.

Whether Currie's account can provide insight into the effect of conservatism on inquiry will therefore require a more sophisticated understanding of creativity. Such an understanding would need to provide a reliable means of picking out those situations wherein the benefits of creativity are not to do with revolutionary theorizing. In those situations, Currie's account could give us some grasp of how to evaluate the epistemic health of the relevant community. But the grounds for that evaluation would ultimately reside in the more sophisticated account. This is because only according to that more sophisticated account could we explain in virtue of what revolutionary theorizing is, in the particular case at hand, rendered unimportant.

5 Conclusion

I have argued that Rovelli's remarks ultimately uncover a shortcoming of Currie's account of creativity. This shortcoming concerns the possible value of revolutionary theorizing to the aims of a research program. Lacking a more sophisticated account of creativity, it is difficult to assess a variety of claims of independent interest. For instance, what commitments does Rovelli make about the problem of quantum gravity, in order to

claim that revolutionary theorizing is valuable within contemporary fundamental physics? And when is it appropriate to focus questions about creativity exclusively on just what is conceived as worth exploring? After all, Currie is unequivocal about the relevance of his more narrow account of creativity in the case of X-risk. He states: "...it is this kind of creativity which scientific study of existential risk requires" [p. 8]. So, by what reasons do the local details of X-risk entitle us to restrict our study to an account that disregards the possibility that revolutionary theorizing matters?

Currie anticipates the possibility that a more sophisticated notion of creativity might ultimately be demanded. By his reckoning, this is because his account does not capture 'ingenuity' (p. 8), failing to distinguish creative searches from chaotic ones. Currie then suggests that a new account of creativity, built on the notion of creative 'flair' developed by Gaut (2010), might capture such a distinction.

This suggestion strikes me as promising. For instance, creative searches might be those hot searches that enable the community to subsequently achieve novelty in research (e.g. at the end of some iterative process). But I would like to conclude by noting one major obstruction to developing the suggestion further. Following Currie, the first step in articulating an account of creativity would be to specify how to extrapolate from the individual to the community level. Such a move is essential to an understanding of the relationship between the social structure of science and creativity, like we have understood it here. (Of particular interest is whether conservatism can be responsible for reliably depriving us of access to much of a solution space, within the developed account.) But extrapolating from the individual to the community level is no small challenge. Creative flair is an irreducibly agential notion, concerning an individual's familiarity with their own goals. It is unclear at present what would mark a

community that, as a whole, is creative in this refined, goal-sensitive respect.

There is, it seems, still much work to be done.

References

Currie, A. (2019). Existential risk, creativity & well-adapted science. *Studies in History and Philosophy of Science Part A 76*, 39–48.

Dardashti, R. (2019). Physics without experiments? In *Why Trust a Theory?: Epistemology of Fundamental Physics*, pp. 154–172. Cambridge University Press.

Gaut, B. (2010). The philosophy of creativity. *Philosophy Compass 5*(12), 1034–1046.

Rovelli, C. (2018). Physics needs philosophy. Philosophy needs physics. *Foundations of Physics 48*(5), 481–491.

Schneider, M. D. (2020). What's the problem with the cosmological constant? *Philosophy of Science 87*(1), 1–20.

Stanford, P. K. (2019). Unconceived alternatives and conservatism in science: the impact of professionalization, peer-review, and Big Science. *Synthese 196*(10), 3915–3932.

Title: Tacking by Conjunction, Genuine Confirmation and Bayesian Convergence

Author: Gerhard Schurz (DCLPS, HHU)

Abstract:

Tacking by conjunction is a well-known problem for Bayesian confirmation theory. In the first section of the paper we point out disadvantages of orthodox Bayesian solution proposals to this problem and develop an alternative solution based on a strengthened concept of probabilistic confirmation, called genuine confirmation. In the second section we illustrate the application of the concept of genuine confirmation to Goodman-type counter-inductive generalizations and to post-facto speculations. In the final section we demonstrate that genuine confirmation is a necessary condition for Bayesian convergence to certainty based on the accumulation of conditionally independent pieces of evidence.

1. From Tacking by Conjunction To Genuine Confirmation

Tacking by conjunction is a deep problem of orthodox Bayesian confirmation theory. It is based on the insight that to each hypothesis H that is confirmed by a piece of evidence E one can 'tack' an irrelevant hypothesis X so that H∧X is also confirmed by

E, in the Bayesian sense of "confirmation" as *probability-raising*, i.e. P(H|E) > P(H) ("P" for "probability"). To illustrate, according to the orthodox account each piece of evidence that confirms Newtonian mechanics also confirms the conjunction of Newtonian mechanics and creationism, although creationism is irrelevant to both Newtonian mechanics and the given evidence. This does not accord well with the pre-theoretic notion of confirmation that Bayesians purport to explicate.

Particularly counterintuitive is the *special* case of tacking by conjunction in which the irrelevant hypothesis is directly tacked to the evidence. Thus E confirms E∧X for every arbitrary hypothesis X, provided only that E and E∧X are *P-contingent*, where a proposition is called "P-contingent" if its probability is different from 0 and 1. For example, "snow is white" confirms "snow is white and creationism". Author (2014) calls this type of 'confirmation' "pseudo-confirmation". The probabilistic fact underlying pseudo-confirmation is simple (Proof in appendix A1):

Theorem 1 (Fact underlying pseudo-confirmation):
Assume H and E are P-contingent. Then E confirms H iff P(E|H) > P(E). Subcase: E |== H. Special case: H = E∧X.

Recent years have seen an increasing interest in the tacking by conjunction problem. Existing Bayesian solution proposals try to soften the negative impact of this result by showing that although H∧X is confirmed by E, it is so only to a lower degree (cf.

Fitelson 2002; Hawthorne and Fitelson 2004, and Crupi and Tentori 2010 who extended the focus to cases where H is disconfirmed by the evidence). Although these solution proposals provide important insights to the Bayesian confirmation model, they suffer from two drawbacks:

(1.) In application to the special case of the tacking problem in which X is directly tacked to E one would intuitively expect the tacked-on hypothesis "E∧X" to not be confirmed at all, but it counts as confirmed according to 'diminished confirmation' proposals.

(2.) These proposals are measure-sensitive in the sense that the 'diminished confirmation' claim holds only for some of the prominent Bayesian confirmation measures, but is violated for others (cf. co-author and author 2019).

One can easily see, however, that E increases the probability of E∧X only because E is a content element of E∧X and increases its own probability to 1 (P(E|E) = 1), while E does not increase the probability of the content element X that *logically transcend*s E, which means by definition that X is not entailed by E. More generally speaking, E does not need to raise the probability of the E-transcending content elements of a hypothesis H, in order to confirm H in the Bayesian sense. Gemes and Earman (Earman 1992, 98n5) have called this type of pseudo-confirmation "confirmation by (mere) *content-cutting*". To avoid this problem one ought to require that the confirmation takes place in those content elements of the hypothesis that are not logically contained in the evidence. Thus, in order for E to count as genuine confirmation of E∧X, E has to confirm X. This is the idea of genuine confirmation devel-

oped in author (2014a) and co-author and author (2019).

The notion of genuine confirmation is based on the notion of a *content element*. A definition of this notion for predicate languages has been given in co-author and author (2017, def. 4.2) and Author 2014b, def. 3.12-2) as follows (where propositional variables count as 0-placed predicates):

Definition 1: C is a content element of (hypothesis) H iff (i) H logically entails C (H |== C), (ii) no predicate in C is replaceable by an arbitrary new predicate with the same place number, salva validitate of H |== C, and (iii) C is elementary in the sense that C is not L(ogically) equivalent with a conjunction $C_1 \wedge C_2$ of conjuncts both of which are *shorter* than C.

The shortness criterion is related to the well-known concept of *minimal description length* in machine learning (Grünwald 2000); it is relativized to an underlying language with $\neg$, $\wedge$, $\vee$,$\exists$ and $\forall$ as primitive logical symbols, assuming that defined symbols are eliminated by their definitions. In propositional logic an equivalent version of this definition has been given in terms of shortest clauses (co-author and author 2017, def. 4.1; 2019, def. 3). Note that $(p \vee q) \wedge (p \vee \neg q)$ is not an admissible conjunctive decomposition of p, which avoids the Popper-Miller (1983) objection to inductive confirmation, which runs as follows: every hypothesis H is logically equivalent to the conjunction $(H \vee E) \wedge (H \vee \neg E)$. But $H \vee E$ is entailed by E and $H \vee \neg E$ is provably

disconfirmed by E, so "inductive" confirmation is impossible. But neither (H∨E) nor (H∨¬E) are content elements of H.

Other technical definitions of content elements are possible – examples are Friedman's (1974) "independently acceptable elements", Gemes' (1994) "content parts" and Fine's (2017) "verifiers". The technical details don't matter as long as the core idea is captured, namely the decomposition of a hypothesis into a set of smallest content elements that are not further conjunctively decomposable in relevant ways and whose conjunction is L-equivalent to the original hypothesis.

The notion of genuine confirmation (GC) has been explicated by co-author and author (2019) in three versions: qualitative full GC, qualitative partial GC and quantitative GC:

Definition 2: Assume E does not entail H.[1] Then:

1.1 Qualitative full GC: E fully genuinely confirms H iff (i) P(X|E) > P(X) holds for all E-transcending content elements X of H.

1.2 Qualitative partial GC: E partially genuinely confirms H iff P(X|E) ≥ P(X) holds for all and P(X|E) > P(X) holds for some E-transcending content elements X of H.

1.3 Quantitative GC: The degree of genuine confirmation that E provides for H is the

[1] We leave it open whether one wants to count logical entailment (E |== H) as a case of 'genuine confirmation' or not. In this case, H has no E-transcending content elements.

sum of the confirmation degrees, conf(E,H), over all E-transcending content elements X of H, divided by their number (where "conf(E,H)" is one of the standard Bayesian confirmation measures, e.g., the difference measure).

Note that although the notion of genuine confirmation (in particular that of genuine full confirmation) strengthens ordinary Bayesian confirmation considerably, it is spelled out within the ordinary Bayesian framework.

2. Applications of Genuine Confirmation

In co-author and author (2019) it is shown that the so-defined measure has a number of attractive features. For example, it can solve problem of measure sensitivity. Moreover, qualitative partial GC implies positive quantitative GC; thus the qualitative and the quantitative notions of GC are in coherence. In this paper we elaborate some attractive features of qualitative confirmation.

Partial (qualitative) genuine confirmation is sufficient to rule out the special case of tacking by conjunction in which the irrelevant hypothesis X is directly tacked on the evidence. This includes an important subcase, namely the problem of Bayesian pseudo-confirmation of Goodman-type *counter-inductive generalizations*. Let E be the evidence that all observed emeralds have been green, H_1* the hypothesis that all unobserved emeralds will be green and H_2* the hypothesis that call unobserved emeralds will be red. Then the inductive generalization H_1 is L-equivalent with $E \wedge H_1$* and

the counter-inductive generalization H_2 is L-equivalent with $E \wedge H_2^*$. Now, following from theorem 1, E confirms both H_1 and H_2 in the pseudo-sense. However, E's confirmation of H_2 is not a genuine one, because E does not confirm H_2's E-transcending content element H_2^*. Moreover, note that E will only confirm the E-transcending inductive projection H_1^* of E, and thus genuinely confirm H_1, if the underlying probability function P satisfies certain additional *inductive* principles, such as de Finetti's exchangeability (invariance of P under permutation of individual constants) and regularity ($P(S) \neq 0, 1$ for every analytically contingent S).

For ruling out all sorts of tacking by conjunction, full (qualitative) genuine confirmation is needed. A further important application of full GC is the elimination of the pseudo-confirmation of *post-facto speculations*. By this we mean the confirmation of hypotheses that contain *theoretical* concepts or, more generally, *latent variables* that are not present in the evidence. By postulating sufficiently many latent variables and suitable principles connecting them with the observed variables, one can explain any observation whatsoever. For example, the fact that grass is green (E) pseudo-confirms the hypothesis (H) that "God wanted that grass is green and whatever God wants, happens". Here "God's wishes" figure as the latent variable. Author (2014a) suggests to understand the pseudo-confirmation of post-facto speculations based on Worrall's (2016) concept of *use-novel evidence*. Worrall's account starts from the observation that the values of the latent variables of a general type of hypothesis are *fitted* towards the evidence. Author (2014a) argues that the unfitted hypothesis H_{unfit} should be understood as a content element of the fitted hypothesis H_{fit},

which is obtained as the existential quantification over the possible values of the latent variables. If H_{unfit} is so general that it can be fitted to every evidence, then H_{unfit} cannot be said to be confirmed merely by the fact that H_{unfit} was fitted to a particular evidence E_1, leading to H_{fit} (although by theorem 1 H_{fit}'s probability has increased, $P(H_{fit}|E_1) > P(H_{fit})$). For example, in the case of the "God-has-wanted-it" hypothesis, H_{unfit} would be the hypothesis "$\exists$X(God wants X and whatever God wants, happens)". According to our account, this hypotheses cannot be genuinely confirmed by theological post-facto explanations of events. This follows straightforwardly from $P(E_1|H_{unfit}) = P(\neg E_1|H_{unfit})$, which holds because H_{unfit} can be fitted to any evidence whatsoever. Only if the fitted hypotheses is confirmed by a second *use-novel* piece of evidence E_2, i.e. one to which H_{unfit} has not been fitted and which H_{fit} could have predicted, then H_{unfit} can be said to be confirmed via the confirmation of H_{fit} by E_1 and E_2. For obviously it is not possible to fit H_{unfit} to a given evidence E_1 and then to confirm the so-obtained H_{fit} by any other evidence E_2 whatsoever. In this way, the concept of genuine confirmation provides a probabilistic justification of Worrall's criterion of use novelty. As a side remark we mention that the use-novelty criterion is by no means a purely philosophical invention, but is employed in a famous computational learning method, namely cross validation (Shalev-Shwartz and Ben-David 2014, sec. 11.2).

When we argued above that the probability of an E-transcending content element of H is or is not raised conditional on an evidence E that raises H's probability, we frequently argued by considerations of intuition. Probability theory itself does not tell us the value of P(E|C). Based on the considerations above we suggest the following

rationality criteria for the spread of the evidence-induced probability increase from a hypothesis H to its E-transcending content elements.:

Necessary criteria for spread of probability increase:

If H increases E's probability, then the resulting probability increase of H by E spreads from H to an E-transcending content element C of H ($P(C|E) > P(C)$) *only if:*

(1.) C is necessary within H to make E probable, i.e., there exists no conjunction H* of content elements of H that makes E at least equally probable ($P(E|H^*) \geq P(E|H)$) but does not entail C, and

(2.) it is not the case that C is an existential quantification, $C = \exists x H(x)$, and H results from a parameter-adjustment of x in H(x) towards the evidence E, such that an equally good fitting of H(x) would have been possible for every possible alternative evidence E'.

In the next section be explain a particular important application of the concept of genuine confirmation: it is a precondition for an important form of Bayesian convergence.

3. From Genuine Confirmation to Bayesian Convergence

An important part of Bayesian epistemology are convergence theorems. According

to them the conditional probability of a hypotheses can be driven to near certainty, if many confirming and mutually conditionally independent pieces of evidence for this hypotheses are accumulated (Earman 1992, 141ff.). Most versions of Bayesian convergence theorems have been formulated for hypotheses not containing latent variables, typically hypotheses that are obtainable from the evidence by enumerative induction. For example, it has been shown that if P is countably additive, then $\lim_{n\to\infty}P(p(Fx)=r \mid (E_1,\ldots,E_n)) = 1$, where each E_i is Fa_i or $\neg Fa_i$ and F's frequency limit in the sequence $(E_1,\ldots,E_n)$ is r (this is a consequence of the theorem of Gaifman and Snir 1982). More important, however, is convergence theorem for hypotheses containing latent variables. A well-known convergence theorem for this case is the following (proof in appendix A2):

Theorem 2 - convergence to certainty:

If a P-contingent hypothesis H satisfies the following conditions

(a) H is confirmed by each of the P-contingent pieces of evidence $E_1,\ldots,E_n$ (i.e., $P(E_i|H) > P(E_i)$ for all $i\in\{1,\ldots,n\}$),

(b) the pieces of evidences are mutually independent conditional on H, i.e., $P(E_i|H\wedge E_1\wedge\ldots\wedge E_{i-1}) = P(E_i|H)$ for all $i \in\{1,\ldots,n\}$ (and some ordering of the E_i's),

(c) and they are also mutually independent conditional on $\neg H$,

then $\lim_{n\to\infty}P(H|E_1\wedge\ldots\wedge E_n) = 1$.

Convergence to certainty in spite of a small prior probability is the ideal case of scientific confirmation. The confirmation of Darwinian evolution theory by multiple pieces of evidence constitutes an example. Theorem 2 is a reformulation of the *Condorcet jury theorem*, with the agreeing reports of the independent witnesses being equated with the independent evidences (Bovens and Hartmann 2003; List 2004). Surprisingly, however, a necessary condition for convergence to certainty is full genuine confirmation. The existence of only one E-transcending content element of H, call if C, that is not confirmed by any one of the evidences E_i, is sufficient to prevent convergence to certainty. Since C's probability is not raised by any of the E_i it holds that $P(C|E_1\wedge\ldots\wedge E_n) = P(C)$. But $P(C|E_1\wedge\ldots\wedge E_n) = P(C)$ is an upper bound of $P(H|E_1\wedge\ldots\wedge E_n)$, since H entails C. Thus $P(H|E_1\wedge\ldots\wedge E_n)$ is forced to stay below P(C), which is small, and cannot approach certainty.

Theorem 3 – failure of convergence to certainty:

If a hypotheses H satisfies conditions (a) and (b) of theorem 2, but contains a content element C that is not confirmed by any of the evidences E_i, then

(i) $\lim_{n\to\infty}P(H|E_1\wedge\ldots\wedge E_n) \leq P(C)$, and

(ii) condition (c) of theorem 2 fails.

Note that if case of theorem 3(i) obtains and H starts from a low prior, then H's probability is still increasing conditional on the accumulating pieces evidence, how-

ever, it does not converge to 1, but to P(C) (from below).

In conclusion, genuine confirmation is a precondition for the sustainable confirmation of hypotheses that are allowed to contain latent variables. While the proof of theorem (i) is obvious from the arguments above, it is prima facie puzzling how this result squares with theorem 2. It turns out that entailment of an irrelevant content elements undermines the independence of the pieces of evidence conditional on the negation of the hypothesis, which is the content of theorem 3(ii). Theorem 3(ii) points towards a general limitation of the convergence theorem 3; because of its importance we state the proof right here in the text (not in the appendix). For whenever the negation of the hypotheses, $\neg H$, can be decomposed into a partition of finer hypotheses that convey different probabilities to the evidence, then the independence of the pieces of evidence conditional on $\neg H$ fails. For example, assume $\neg H$ splits into two disjoint hypotheses H_2, H_3 such that $P(E_i|H_2)$ is much larger than $P(E_i|H_3)$ (for all i), although $(E_i|H_2 \vee H_3) = P(E_i|\neg H) < P(E_i)$, which follows from $P(E_i) < P(E_i|H)$ and the P-contingency of E_i and H. Then $P(E_j|\neg H \wedge E_i) > P(E_j|\neg H)$ will hold, because the fact that E_i obtained makes it more probable that H_2 and not H_3 obtained, which in turn makes E_j more probable.

Now assume that H is a hypothesis that has an irrelevant content element C, $H = H_1 \wedge C$, where $P(E_i|H_1) > P(E_i)$ and C is irrelevant for E_i both unconditionally and conditionally on H_1. In this case the negation $\neg(H_1 \wedge C)$ splits into the finer partition $\neg H_1 \wedge C$, $\neg H_1 \wedge \neg C$ and $H_1 \wedge C$. While $P(E_i|\neg H_1 \wedge \pm C) < P(E_i)$ holds for both $\pm C = C$ and

$\pm C = \neg C$, the third element of the partition behaves differently, namely $P(E_i|H_1 \wedge \neg C) > P(E_i|H_1)$, and this destroys the independence of the evidence conditional on $\neg(H_1 \wedge C)$.

Fortunately there is a generalized version of theorem 3 that is relativized to a given possibly large partition of hypotheses that are assumed to be sufficiently strong to guarantee mutual conditional independence of the pieces of evidence (proof in appendix A3):

Theorem 4 - generalized convergence to certainty:

Assume a P-contingent hypothesis H_1 belongs to a partition of hypotheses $\{H_1,\ldots,H_m\}$ satisfying the following conditions:

(a) every piece of evidence *favors* H_1 over every other hypothesis by at least δ (for some $\delta>0$), i.e., $P(E_i|H_1) \geq P(E_i|H_r)+\delta$ for all $r>1$ and $i \in \{1,\ldots,n\}$, and

(b) the pieces of evidences are mutually independent conditional on every H_k ($k \in \{1,\ldots,m\}$), i.e., $P(E_i|H_k \wedge E_1 \wedge \ldots \wedge E_{i-1}) = P(E_i|H)$ for all i ($i \in \{1,\ldots,n\}$),

then (i) $P(H_1|E_1 \wedge \ldots \wedge E_n) \geq \frac{h}{h + (1-h)\cdot(1-\delta)^n}$, and

(ii) $\lim_{n\to\infty} P(H_1|E_1 \wedge \ldots \wedge E_n) = 1$.

If we apply theorem 4 to hypotheses that are conjunctions of several content elements, $H = H_1 \wedge \ldots \wedge H_k$, then the smallest partition of competing hypotheses that has to be checked in regard to conditional independence of the pieces of evidence is the

partition$\{\pm H_1\wedge\ldots\wedge\pm H_k: \pm H_i \in \{H_i,\neg H_i\}, 1\leq i\leq k\}$, which contains 2^k elements.

Appendix: Proof of theorems:

A1. Proof of theorem 1:

This is well-known: Assuming H and E are P-contingent, then

$P(H|E) = P(H)\cdot P(E|H)/P(E)$, and $P(E|H)\cdot P(H)/P(E) > P(H)$ iff $P(E|H) > P(E)$. Q.E.D.

A2. Proof of theorem 2:

Theorem 2 follows from theorem 4 by substituting $\{H,\neg H\}$ for $\{H_1,\ldots,H_m\}$. Note that for P-contingent E and H, $P(E|H) > P(E)$ entails $P(E) > P(E|\neg H)$, which follows from the fact that $P(E) = P(E|H)\cdot P(H) + P(E|\neg H)\cdot P(\neg H)$. Thus there exists a δ such that $P(E|H) \geq P(E|\neg H)+\delta$, which is the assumption of theorem 3. Q.E.D.

A3. Proof of theorem 4:

We abbreviate $P(E_i|H_1)$ as p_i and write $\Sigma\{x_1,\ldots,x_n\}$ and $\Pi\{x_1,\ldots,x_n\}$ for the sum and the product of the numbers $x_1,\ldots,x_n$, respectively. We calculate as follows. By Bayes' theorem:

$$P(H_1|E_1\wedge\ldots\wedge E_n) = P(E_1\wedge\ldots\wedge E_n|H_1)\cdot P(H_1)/\Sigma\{P(E_1\wedge\ldots\wedge E_n|H_r)\cdot P(H_r): 1 < r \leq m\}.$$

Since $P(E_1\wedge\ldots\wedge E_n|H_r) = \Pi\{P(E_i|H_r\wedge E_1\wedge\ldots\wedge E_{i-1}): 1\leq i\leq n\}$ and condition (b) of theorem 4 we continue:

$$= \frac{h \cdot \Pi\{p_i : 1 \le i \le n\}}{h \cdot \Pi\{p_i : 1 \le i \le n\} + \Sigma\{P(H_r) \cdot \Pi\{P(E_i|H_r) : 1 \le i \le n\} : 1 \le r \le m, r > 1\}}$$

$$\geq \frac{h \cdot \Pi\{p_i : 1 \le i \le n\}}{h \cdot \Pi\{p_i : 1 \le i \le n\} + \Sigma\{P(H_r) : 1 \le r \le m, r > 1\} \cdot \Pi\{(p_i - \delta) : 1 \le i \le n\}} \text{ (from condition (a))}$$

$$= \frac{h \cdot \Pi\{p_i : 1 \le i \le n\}}{h \cdot \Pi\{p_i : 1 \le i \le n\} + (1-h) \cdot \Pi\{(p_i - \delta) : 1 \le i \le n\}} = \frac{1}{1 + \frac{1-h}{h} \cdot \frac{\Pi\{(p_i - \delta) : 1 \le i \le n\}}{\Pi\{p_i : 1 \le i \le n\}}} .$$

Because of $\frac{\Pi\{(p_i - \delta) : 1 \le i \le n\}}{\Pi\{p_i : 1 \le i \le n\}} \le (1-\delta)^n$ we obtain the claim of theorem 4 (i), which entails theorem 4 (ii) because of $\lim_{n\to\infty}(1-\delta)^n = 0$. Q.E.D.

References (blinded)

Author (2014a). in *Studies in History and Philosophy of Science*.

Author (2014b). [Book]

Bovens, L., and Hartmann, S. (2003). *Bayesian Epistemology*, Oxford University Press, Oxford.

Co-author and author (2019). in *Studia Logica*.

Co-author and author (2019). in *British Journal for the Philosophy of Science*.

Crupi, V. and Tentori, K. (2010). "Irrelevant Conjunction: Statement and Solution of a New Paradox", *Philosophy of Science*, 77, pp. 1–13.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, Cambridge, MA: MIT Press.

Fitelson, B. (2002). "Putting the Irrelevance Back into the Problem of Irrelevant Con-

junction", *Philosophy of Science*, 69, 611–22.

Gaifman, H., and Snir, M. (1982). "Probabilities Over Rich Languages", *Journal of Symbolic Logic*, 47, 495-548.

Grünwald, P. (2000). "Model Selection Based on Minimal Description Length", *Journal of Mathematical Psychology*, 44, 133-152.

Hawthorne, J. and Fitelson, B. (2004). "Discussion: Resolving Irrelevant Conjunction with Probabilistic Independence", *Philosophy of Science*, 71, 505–14.

List, C. (2004). "On the Significance of the Absolute Margin", *British Journal for the Philosophy of Science*, 55, pp. 521–44.

Popper, K., and Miller, D. (1983). "A Poof of the Impossibility of Inductive Probability", *Nature*, 302, 687-688.

Shalev-Shwartz, S. and S. Ben-David. (2014). *Understanding Machine Learning. From Theory to Algorithms*, Cambridge University Press, New York.

entropy

MDPI

Article

Beyond Causal Explanation: Einstein's Principle Not Reichenbach's

Michael Silberstein [1,2,*], William Mark Stuckey [3] and Timothy McDevitt [4]

[1] Department of Philosophy, Elizabethtown College, Elizabethtown, PA 17022, USA
[2] Department of Philosophy, University of Maryland, College Park, MD 20742, USA
[3] Department of Physics, Elizabethtown College, Elizabethtown, PA 17022, USA; stuckeym@etown.edu
[4] Department of Mathematical Sciences, Elizabethtown College, Elizabethtown, PA 17022, USA; mcdevittt@etown.edu
* Correspondence: silbermd@etown.edu

Abstract: Our account provides a local, realist and fully non-causal principle explanation for EPR correlations, contextuality, no-signalling, and the Tsirelson bound. Indeed, the account herein is fully consistent with the causal structure of Minkowski spacetime. We argue that retrocausal accounts of quantum mechanics are problematic precisely because they do not fully transcend the assumption that causal or constructive explanation must always be fundamental. Unlike retrocausal accounts, our principle explanation is a complete rejection of Reichenbach's Principle. Furthermore, we will argue that the basis for our principle account of quantum mechanics is the physical principle sought by quantum information theorists for their reconstructions of quantum mechanics. Finally, we explain why our account is both fully realist and psi-epistemic.

Keywords: EPR correlations; relativity principle; principle explanation; Reichenbach's Principle; retrocausality; locality; contextuality; no preferred reference frame; causal modelling; no-signalling; quantum information theory; reconstructions of quantum mechanics; Tsirelson bound; realist psi-epistemic

check for updates

Citation: Silberstein, M.; Stuckey, W.M.; McDevitt, T. Beyond Causal Explanation: Einstein's Principle Not Reichenbach's. *Entropy* **2021**, *23*, 114. https://doi.org/10.3390/e23010114

Received: 16 December 2020
Accepted: 11 January 2021
Published: 16 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

There is a class of interpretations or accounts of quantum mechanics (QM) called retrocausal theories (for more historical background and comparisons of different models, see [1,2]). Such models vary wildly, and it would seem that the only thing they have in common is that the future determines the past or present as much as the past or present determines the future, at least with respect to some QM phenomena. However, many of the purveyors of retrocausal accounts do have similar motives. Namely, to show that QM does not, contrary to certain "no-go theorems," entail non-locality, contextuality and realism about the wavefunction and QM states. Furthermore, defenders generally agree that a retrocausal account ought to be nonetheless a realist account of QM. It is for this reason that we cannot avoid delving into some detail on the question of what constitutes a realist account of QM. Furthermore, the discussion of this topic will come in handy in explaining why our account is realist and psi-epistemic.

What exactly makes an interpretation of QM a realist one is up for debate, but at the very least, it is generally believed that realist interpretations cannot be purely epistemic. For example, according to QBism, QM probabilities are not about objective reality, rather they are about updating the belief states of single epistemic agents. Healey's pragmatist account of QM [3] and Rovelli's relational account of QM [4] both hold that the QM state is not a description of the physical world, but only exists to generate QM probabilities. Unlike QBism, both pragmatist accounts and relational accounts of QM are relative-state theories in a sense, the difference is that in the pragmatist account a quantum state ascription is relative only to the perspective of an actual or potential agent, whereas in relational QM values are

objective and relative to any physical system—information is relative information that one physical system has about another, as with a third physical system observing two other entangled systems, etc.

The point is that in relational QM, all information is purely relational and this need have nothing to do with 'agents' even in the neutral sense of agent in which a non-conscious A.I. might be observing and measuring outcomes in QM experiments. However, Healey is clear that a QM state can only be ascribed to an agent in the context of an experimental set-up that defines the perspective of that agent; in this respect Healey's pragmatist QM is a kind of half-way house between QBism and relational QM. In spite of the agent-centric talk in QBism and the pragmatist account, these accounts do not require *conscious* agents. An epistemic agent could be a non-conscious machine of some sort. What makes all three of these accounts epistemic is that they all hold the QM state is not a description of the physical world, but only exists to generate QM probabilities. That is, in addition to being explicitly psi-epistemic, none of these accounts provides an ontology or what Bell called [5] "beables," that are allegedly hiding behind the veil of the 'observables' and explain the phenomenology in question. The beables, such as particles, fields or waves of some sort, are supposed to tell us exactly what happens, say, between the initiation and termination of some QM experiment, such as a Bell-type experiment or twin-slit type experiment. Clearly this notion of beables presupposes what we will shortly define as a dynamical or causal explanatory bias.

It is sometimes further claimed that beables must have some metaphysical autonomy/independence and some intrinsic properties. If that is so, then relational QM fails to be a realist theory for yet another reason. As Laudisa and Rovelli put it [4]:

> For RQM (relational quantum mechanics), the lesson of quantum theory is that the description of the way distinct physical systems affect each other when they interact (and not the way physical systems 'are') exhausts all that can be said about the physical world. The physical world must be described as a net of interacting components, where there is no meaning to 'the state of an isolated system', or the value of the variables of an isolated system. The state of a physical system is the net of the relations it entertains with the surrounding systems. The physical structure of the world is identified as this net of relationships.

Thus, if relational QM is true, there are no such things as beables so defined. We will return to such questions in the Discussion and Postscript wherein we will take up the topic of contextuality and realism more explicitly. Therein, we will explain why our principle account of QM is a realist, psi-epistemic account, as well as our take on beables, etc.

Finally, in perhaps the most egregious violation of realism, some of these accounts, such as relational QM, are labeled subjectivist because they allegedly entail that, at least in certain situations such as Wigner's friend type set-ups, different observers can consistently give different accounts of the same set of events such as the outcomes of measurements. For example, in a particular Schrödinger's Cat type set-up, even without invoking the branching structure of the Many-Worlds interpretation, observer X can report seeing a live cat and observer Y can report seeing a dead cat, allegedly the very same cat, and both can be correct without contradiction [6] (pp. 116–117). That is, such subjectivist accounts allegedly violate what is sometimes called, The Absoluteness of Observed Events [7]. For a detailed explanation of why our principle account of QM rules out such absurdities see the Postscript and [8].

Many purveyors of retrocausal models of QM are hoping to construct realist models in the sense of being housed strictly in spacetime and requiring nothing but beables such as particles, semi-classical fields or semi-classical waves. Perhaps, then, it is not surprising that most retrocausal accounts, even those that have a block universe picture firmly in mind, still insist that the best explanation for EPR correlations must be a causal explanation of some sort. The idea here being that what it means to provide a realist account of said correlations is to provide a causal account of some sort. Whatever their motivations and whatever their particular account of retrocausation might be, such theorists still adhere

to some version of Reichenbach's Principle, which states that if two events are correlated, then either there is a causal connection between the correlated events that is responsible for the correlation or there is a third event, a so-called common cause, which brings about the correlation.

When it comes to how retrocausal accounts might thwart various no-go theorems, the focus is on the statistical "measurement independence" assumption in Bell's theorem, i.e., the assumption that outcomes in EPR-type experiments do not causally depend on any future measurement settings. If one could construct a fully time-like retrocausal account of QM that fully explained EPR correlations with no remaining fatal flaws and which reproduced the statistics of QM, then perhaps locality (what Bell calls "local causality") could be saved. Furthermore, if one can tell such a retrocausal account, then perhaps contrary to the Kochen–Specker theorem and others like it [9], non-contextuality can be saved as well: the claim that physical properties of QM systems exist prior to and independently of the act of measurement. As Friederich and Evans put it [1]:

> Retrocausality renders Kochen-Specker-type contextuality potentially explainable as a form of "causal contextuality". If there is a backward-directed influence of the chosen measurement setting (and context) on the pre-measurement ontic state, it is no longer to be expected that the measurement process is simply uncovering an independently existing definite value for some property of the system, rather the measurement process can play a causal role in bringing about such values (the measurement process is retrocausal rather than retrodictive). Indeed, one might argue contextuality of measured values is just what one might expect when admitting retrocausal influences. As Wharton (2014: 203) puts it, "Kochen-Specker contextuality is the failure of the Independence Assumption", i.e., the failure of measurement independence.

Finally, the idea is that if retrocausation can thwart non-locality and contextuality, then perhaps it provides the basis for a realist, psi-epistemic account of QM in spacetime alone, with counterfactual definiteness and determinate physical properties throughout the worldtube of every QM system. Of course, while many retrocausal accounts adhere to Reichenbach's Principle in some form, the nature of the causal relation itself varies across different retrocausal explanations. However, retrocausal explanations tend to invoke one of two (or both) notions of causation. The first kind of causation is a "causal processes account," wherein chains of events are related by causal interactions ("action-by-contact"), that involve local exchanges of a conserved quantity. Such causal influences extend through spacetime via contiguously mediated connections between local beables, as with classical fields [10]. The trick for retrocausal accounts espousing this type of causation is to thwart non-locality (faster-than-light causal connections between space-like separated events, i.e., "spooky actions at a distance" [11] (p. 158)), by telling a story whereby such causal processes are purely time-like. Such causal processes are often described as making a time-like "zig-zag" pattern in spacetime between the two space-like separated detectors in a standard EPR-type setup [12–14]. We should note that some such accounts are still realist about the QM wavefunction or at least about semi-classical waves in spacetime (the general idea here is that wavefunctions evolve both forwards and backwards in time), but not all [15,16].

The second type of causation is called the "interventionist" or "manipulability" account of causation [17]. The central idea is that X is a cause of Y if and only if manipulating X is an effective means of indirectly manipulating Y. According to retrocausal accounts of QM espousing an interventionist account of causation, manipulating the setting of a measurement apparatus now can be an effective means of manipulating aspects of the past. The formal machinery of causal modelling has the interventionist account of causality as its foundation [18].

Price and Wharton, two key defenders of retrocausal accounts of QM, embrace a subset of interventionism known as the "agent" or "perspectivalism" account of causation [19–21]. On this view, causal relations are relations that can be used for control or manipulation,

from the perspective of the agent in question of course. This is an understandably appealing notion of causation for those such as Price and Wharton who espouse a block universe picture, wherein causation talk cannot possibly be about changing or bringing about events (past, present or future) in any robust sense of those terms. In our language [22], agent causation focuses on the "ant's-eye" view of explanation from within the block universe, as opposed to the "God's-eye" view that would seek a purely objective explanation for EPR correlations external to a perspective from within the block universe, an explanation that transcends and subsumes perspectival causation, such as conservation laws.

In addition to the specific problems faced by particular retrocausal models, the consensus is that, as of yet, no realist retrocausal account manages to successfully save locality, non-contextuality and a psi-epistemic account of the wavefunction [1]. The most general concern however is that retrocausal accounts fail to provide a robust or coherent causal explanation of EPR correlations and contextuality. The most general form of this concern is that, at least from the God's-eye point of view, the very idea of retrocausation in a block universe makes very little physical or explanatory sense [13,23,24]. Many (though not all [13]) believe that retrocausation demands a block universe. This is because it is hard to see how the future or future boundary conditions could cause anything or participate in any type of explanation of EPR correlations, if the future does not exist. Yet, when we think of the block universe from the God's-eye point of view, it is clear that causation cannot be about bringing new events into being that did not formerly exist, because from a God's-eye point of view it is all just 'there', including EPR-experiments from initiation (source) to termination (detector). The very idea of "causality flowing backwards in time" as with the "causal processes account," simply seems superfluous or redundant in such a world. For example, as Cramer says himself, the backwards-causal elements of his transactional interpretation are "only a pedagogical convention," and that in fact "the process is atemporal" [25] (p. 661). But the idea of an "atemporal process" seems like a non-sequitur. In a block universe, why bother trying to add some new mechanism (such as waves from the future) to account for how information from the future got to the emission event in the past? Again, from a God's-eye point of view the relevant information at every point in the "process" from source to detector, is all just 'there'. Aside from thwarting non-locality, how is this backward brand of causation any better at saving constructive or commonsense notions of causation than "instantaneous causation" between space-like separated events?

Those who advocate for an interventionist or perspectivalist account of causation would argue that such an account of causation still makes sense even in a block universe. However, there are problems with this account of causation as well. As Friederich and Evans note [1]:

> Two of the more significant assumptions are (i) the causal Markov condition, which ensures that every statistical dependence in the data results in a causal dependence in the model—essentially a formalization of Reichenbach's common cause principle—and (ii) faithfulness, which ensures that every statistical independence implies a causal independence, or no causal independence is the result of a fine-tuning of the model.
>
> It has long been recognized (Butterfield 1992; Hausman 1999; Hausman and Woodward 1999) that quantum correlations force one to give up at least one of the assumptions usually made in the causal modeling framework. Wood and Spekkens (2015) argue that any causal model purporting to causally explain the observed quantum correlations must be fine-tuned (i.e., must violate the faithfulness assumption). More precisely, according to them, since the observed statistical independences in an entangled bipartite quantum system imply no signalling between the parties, when it is then assumed that every statistical independence implies a causal independence (which is what faithfulness dictates), it must be inferred that there can be no (direct or mediated) causal link between the parties. Since there is an observed statistical dependence between the outcomes of measurements on the bipartite system, we can no longer account

> for this dependence with a causal link unless this link is fine tuned to ensure that the no-signalling independences still hold. There is thus a fundamental tension between the observed quantum correlations and the no-signalling requirement, the faithfulness assumption and the possibility of a causal explanation.

We would say that even if interventionist and causal modelling accounts of causation could be applied to EPR correlations with nothing like the preceding concerns, there is still little reason to find such explanations deeply satisfying. Is there really no more fundamental and objective, God's-eye explanation for EPR correlations that transcends and subsumes perspectival causation? Such interventionist explanations strike us as too cheap and easy, and not very deep from the perspective of fundamental physics.

In addition to the foregoing concerns, there are recent no-go theorems which allege that no account of QM can escape contextuality, because it is necessary to reproduce the observed statistics of quantum theory [9]. More recent no-go theorems allege to show that not even accounts that give up measurement independence, such as retrocausal, superdeterministic or even non-local models of QM can escape one or another strong form of contextuality, going so far as to claim that what is contextual is not just the QM state, but many other features of QM, such as what counts as a system, dynamical law and boundary conditions [26]. Going even further, Bong et al. [7] allege to provide a new and more powerful no-go theorem that we must give up at least one of the following assumptions [7]:

- Assumption 1 (Absoluteness of Observed Events-AOE): An observed event is a real single event, and not relative to anything or anyone (realism and non-contextuality).
- Assumption 2 (No-Superdeterminism-NSD): Any set of events on a space-like hypersurface is uncorrelated with any set of freely chosen actions subsequent to that space-like hypersurface.
- Assumption 3 (Locality-L): The probability of an observable event e is unchanged by conditioning on a space- like-separated free choice z, even if it is already conditioned on other events not in the future light-cone of z.
- Assumption 4 (The completeness of QM-COMP): QM unmodified applies to any and all macroscopic measuring devices including human observers.

Based on these and similar results, other people make even stronger claims about what the new no-go theorems and experiments show. For example, Renner claims the new theorems are telling us that QM needs to be replaced [27]. Herein, we do not address any of these new no-go theorems directly, we simply note that if these no-go results stand and if the primary goal for most retrocausal accounts is to save locality, non-contextuality, psi-epistemic, and something like classical realism, things are looking increasingly grim.

One might think that advocates of retrocausal accounts would take heart from the Bong et al. results [7], because at least it leaves superdeterminism as an option, and superdeterminism is one way to give up measurement independence. While retrocausal accounts are often labeled as superdeterministic it is important to see that they are different. Technically speaking, in a superdeterministic world, measurement independence is violated via a past common cause, for example, a common cause of one's choice of measurements and the particle spin properties in the case of Bell correlations. Thus, superdeterminism is a conspiratorial theory with only past-to-future causation. It is true that superdeterminism entails that experimenters are not free to choose what to measure without being influenced by events in the distant past, and thus it does give up measurement independence, however, it does so in a particularly spooky way. Superdeterminism forces us to accept some very special conditions at the big bang as a brute fact or seek some sort of physically acceptable explanation for those initial conditions, that is presumably not some sort of supernatural conspiracy. While there are those who defend superdeterminism [28], most retrocausal theorists want to avoid it for the foregoing reasons.

In our book, we noted that most people are predisposed to think dynamically/causally because our perceptions are formed in a time-evolved fashion. Therefore, we want to understand/explain what we experience dynamically/causally [22]. We call this the dynamical or causal explanatory bias. It is not surprising that most people, including

philosophers and physicists, have this bias. What is maybe somewhat surprising is that, as we have just seen, even retrocausal thinkers and blockworlders share this bias, i.e., they embrace the causal processes model and/or the perspectival causation model of explanation as fundamental or essential in some way. Take the following admonition from Price and Wharton [29] (p. 123):

> In putting future and past on an equal footing, this kind of approach is different in spirit from (and quite possibly formally incompatible with) a more familiar style of physics: one in which the past continually generates the future, like a computer running through the steps in an algorithm. However, our usual preference for the computer-like model may simply reflect an anthropocentric bias. It is a good model for creatures like us, who acquire knowledge sequentially, past to future, and hence find it useful to update their predictions in the same way. But there is no guarantee that the principles on which the universe is constructed are of the sort that happens to be useful to creatures in our particular situation. Physics has certainly overcome such biases before—the Earth isn't the center of the universe, our sun is just one of many, there is no preferred frame of reference. Now, perhaps there's one further anthropocentric attitude that needs to go: the idea that the universe is as "in the dark" about the future as we are ourselves.

We share their sentiment, but even leading retrocausalists Price and Wharton are committed to causal explanations of EPR correlations of either the causal processes account or the interventionist/perspectivalist account [1].

As Friederich and Evans suggest and we concur, aside from ourselves, Wharton and Price have come the farthest in moving away from the dynamical/causal explanatory bias. Here is how they describe Wharton's view [1]:

> The account is a retrocausal picture based on Hamilton's principle and the symmetric constraint of both initial and final boundary conditions to construct equations of motion from a Lagrangian, and is a natural setting for a perspectival interventionist account of causality. Wharton treats external measurements as physical constraints imposed on a system in the same way that boundary constraints are imposed on the action integral of Hamilton's principle; the final measurement does not simply reveal preexisting values of the parameters, but constrains those values (just as the initial boundary condition would). Wharton's model has been described as an "all-at-once" approach, since the dynamics of physical systems between an initial and final boundary emerges en bloc as the solution to a two-time boundary value problem.
>
> On this interpretation, one considers reality exclusively between two temporal boundaries as being described by a classical field ϕ that is a solution to the Klein-Gordon equation: specification of field values at both an initial and final boundary (as opposed to field values and their rate of change at only the initial boundary) constrains the field solutions between the boundaries.

While Wharton's "all-at-once" or "Lagrangian" model goes some way toward relinquishing said bias, as noted above, it still falls within the causal processes account and the interventionist/perspectivalist account of causal explanation. After all, one goal of Wharton's retrocausal Lagrangian method is to "fill in" the classical field between initial (source) and final boundary conditions (detector). The Lagrangian method begins describing the space of possible space-time trajectories of the system between two boundary conditions, and then a least action principle such as the path of least time—a global constraint—is used to fix which of these trajectories is actual. More recently Wharton has focused on constructing ignorance-based interpretations of the path integral formalism [15]. The bottom line is that Reichenbach's Principle is still the "axiom of choice" even when it comes to "all-at-once" or "Lagrangian" models of EPR correlations.

As we said, there is as of yet no retrocausal model that recovers the statistics of QM and also saves locality, non-contextuality, psi-epistemic, and classical realism. Over the

years we have argued that the problem with retrocausal accounts is that they do not go far enough in relinquishing their dynamical/causal explanatory bias [2,22,23]. It is not enough for such models to be temporally symmetric, but rather when it comes to EPR correlations we ought to cast off the dynamical and causal mode of explanation completely, in favor of explanation a la adynamical and acausal global constraints (AGC). For example, we need not worry about the traces particles carry of their dynamical interactions, whether past or future. For us, the explanatory requirement of a particle traveling along a determinate trajectory is a holdover from the dynamical/causal bias as exemplified by the causal processes account of explanation. Rather, we should seek a quantum event or action connecting the source and detector. Lewis gets us exactly right in his description of our view as follows [30] (p. 187):

> The idea is to give up the search for forward-acting and backward- acting dynamical laws that can somehow "fit together" in a consistent way to yield quantum phenomena. Rather, we derive quantum phenomena directly from a global constraint, without any appeal to the dynamical evolution of particle properties or wave functions.

We are primarily inspired to construct an AGC-based physical model by the belief that what both relativity and QM are trying to tell us, is that sometimes AGC-type explanations and contextuality are more fundamental than causal or dynamical explanation. In terms of our specific motivations as regards QM, we are primarily interested in constructing a realist, psi-epistemic, local account of QM, that fully comports with a realist view of Minkowski spacetime. That is, we do not believe in Hilbert space/wavefunction realism and we do not believe there is any action-at-a-distance that violates the causal structure (the light-cone structure) of special relativity (SR). Obviously, for us, a realist account of QM need not involve causal or dynamical explanation; we are perhaps alone in fully rejecting Reichenbach's Principle for such cases. However, a realist account of QM ought to explain why EPR correlations exist in Minkowski spacetime. Equally obvious, if we are right in fully rejecting Reichenbach's Principle and the dynamical/causal explanatory bias in QM, then all of these debates about the various problems of retrocausal models are a red herring.

It is one thing to posit explanations in terms of adynamical and acausal global constraints, but it is quite another to cook up a specific model. In our book [22], we posited an ontology of four-dimensional entities that do not move or change but make up the block universe. We also posited an AGC mode of explanation for EPR correlations, among other things, that uses the initial and final states of the system, plus the AGC, to provide a spatiotemporal explanation of said correlations. However, even those somewhat open to our relatively "radical" project of completely jettisoning Reichenbach's Principle and the dynamical and causal explanatory bias behind it, found our specific account formally daunting, vague, and insufficiently precise. Lewis expresses a common concern [30] (p. 188):

> But the way is not altogether clear. Classical adynamical techniques, such as least-action calculations, output a determinate trajectory between two points. But quantum adynamical techniques, such as Feynman's path-integral calculation, output a probability value based on a sum over all possible trajectories between the two points. Which trajectory does the particle take? And what does the probability represent?
>
> Silberstein, Stuckey and McDevitt take this situation to point to direct action between the source and the detector: But what of the probability? A global constraint that rules out non-parabolic baseball trajectories is easy to comprehend. But it is harder to figure out how to understand a probabilistic global constraint. What is constrained, exactly? The frequency of this kind of event?

Take the following even more telling reaction to our book [31] (p. 344):

> I am not sold that the adynamical picture is truly explanatory. Philosophers of science have proposed objective accounts of explanation, but they all recognize

> there's a strong sense in which explanation is 'explanation for us,' and any account should capture our intuition that explanation is fundamentally dynamical. This is connected with causation: intuitively, we explain an event because we find its causes; causes happen before their effects and 'bring them about.' An empiricist will be skeptical of causation, like presumably SSM. However, as is well known, one can dispense of causation and propose models of explanations in which laws of nature and unification of phenomena play an important role. Should I think of SSM's adynamical view in this sense? Or should I connect their view with the distinction between constructive and principle theories, proposed by Einstein (1919)? According to Einstein, principle theories (like thermodynamics) are formulated in terms of principles that systematize the phenomena; so that one has explained an event if it follows from the principles. In contrast, in a constructive theory (as kinetic theory) a phenomenon is explained when it fits into the 'mechanical' model of the theory. Should I understand SSM's view as a principle theory? (But if so, which are the principles?).

In the preceding passage, Allori beautifully expresses the aforementioned recalcitrance of the dynamical and causal explanatory bias. However, more importantly, Allori suggests another way to conceive of our project in terms of providing a principle versus constructive account of QM generally and EPR correlations specifically. This is precisely what we have done in recent subsequent work [32–34], and we will expand upon those results herein.

Finally, our principle account of QM introduced in Section 3 shows a profound unity between QM and SR that is generally unappreciated, especially since by "QM" we are referring to non-relativistic quantum mechanics. We begin in Section 2, with an overview of principle versus constructive explanation in general and the recent history of that debate within QM itself. We present our principle account of QM in Section 3, showing how it resolves a number of QM mysteries. We conclude with Section 4, where we defend our principle account of QM and its obvious implication for causality in physics. In the Postscript we will return to the question of why our principle account is both realist and psi-epistemic, the place of contextuality, etc.

2. Principle Versus Constructive Explanation

Here, we begin with some background needed to appreciate our explanatory project. As we will see, some theorists in QM, such as Fuchs and Hardy, point to the postulates of SR as an example of what quantum information theorists (QIT) seek for QM, and SR is a "principle theory" [35]. That is, the postulates of SR are constraints without a corresponding "constructive" or causal explanation. Here, Einstein explains the difference between the two [36]:

> We can distinguish various kinds of theories in physics. Most of them are constructive. They attempt to build up a picture of the more complex phenomena out of the materials of a relatively simple formal scheme from which they start out. [Statistical mechanics is an example.] ...
>
> Along with this most important class of theories there exists a second, which I will call "principle-theories." These employ the analytic, not the synthetic, method. The elements which form their basis and starting point are not hypothetically constructed but empirically discovered ones, general characteristics of natural processes, principles that give rise to mathematically formulated criteria which the separate processes or the theoretical representations of them have to satisfy. [Thermodynamics is an example.] ...
>
> The advantages of the constructive theory are completeness, adaptability, and clearness, those of the principle theory are logical perfection and security of the foundations. The theory of relativity belongs to the latter class.

Concerning his decision to produce a principle theory instead of a constructive theory of SR, Einstein writes [37] (pp. 51–52):

> By and by I despaired of the possibility of discovering the true laws by means of constructive efforts based on known facts. The longer and the more despairingly I tried, the more I came to the conviction that only the discovery of a universal formal principle could lead us to assured results.

That is, "there is no mention in relativity of exactly *how* clocks slow, or *why* meter sticks shrink" (no "constructive efforts"), nonetheless the principles of SR are so compelling that "physicists always seem so sure about the particular theory of Special Relativity, when so many others have been superseded in the meantime" [38].

Today, we find ourselves in a similar situation with QM. That is, 85 years after the famous EPR paper [39] we still have no consensus constructive account of QM. This prompted Smolin to write [40] (p. 227):

> So, my conclusion is that we need to back off from our models, postpone conjectures about constituents, and begin talking about principles.

Fuchs writes [41] (p. 285):

> Compare [quantum mechanics] to one of our other great physical theories, special relativity. One could make the statement of it in terms of some very crisp and clear physical principles: The speed of light is constant in all inertial frames, and the laws of physics are the same in all inertial frames. And it struck me that if we couldn't take the structure of quantum theory and change it from this very overt mathematical speak ..., then the debate would go on forever and ever. And it seemed like a worthwhile exercise to try to reduce the mathematical structure of quantum mechanics to some crisp physical statements.

And, Hardy writes [42]:

> The standard axioms of [quantum theory] are rather ad hoc. Where does this structure come from? Can we write down natural axioms, principles, laws, or postulates from which we can derive this structure? Compare with the Lorentz transformations and Einstein's two postulates for special relativity. Or compare with Kepler's Laws and Newton's Laws. The standard axioms of quantum theory look rather ad hoc like the Lorentz transformations or Kepler's laws. Can we find a natural set of postulates for quantum theory that are akin to Einstein's or Newton's laws?

Along those lines, QIT have produced several reconstructions of QM, but they are so far not compelling. Dakic and Brukner write [43] :

> The vast majority of attempts to find physical principles behind quantum theory either fail to single out the theory uniquely or are based on highly abstract mathematical assumptions without an immediate physical meaning (e.g., [18]).
> ...
> While [the instrumentalist] reconstructions are based on a short set of simple axioms, they still partially use mathematical language in their formulation. ...
> It is clear from the previous discussion that the question on basis of which physical principles quantum theory can be separated from the multitude of possible generalized probability theories is still open.

Another problem with the reconstructions of QIT is noted by Van Camp [44]:

> However, nothing additional has been shown to be incorporated into an information-theoretic reformulation of QM beyond what is contained in QM itself. It is hard to see how it could offer more unification of the phenomena than QM already does since they are equivalent, and so it is not offering any explanatory value on this front.

Moreover, Fuchs quotes Wheeler, "If one really understood the central point and its necessity in the construction of the world, one ought to state it in one clear, simple sentence" [41] (p. 302). Asked if he had such a sentence, Fuchs responded, "No, that's my big

failure at this point" [41] (p. 302). Herein, we answer the desideratum of QIT explicitly by showing how the relativity principle, aka "no preferred reference frame" (NPRF), is the physical principle corresponding to the reconstructions of QM, just as it is for the Lorentz transformations of SR.

Our claim about principle explanation being "fundamental" deserves some unpacking. Obviously, the question of what makes an explanation relatively "fundamental" is multifaceted (i.e., there are multiple senses of "fundamental") and value laden. Our claim about the fundamentality of principle explanation in this case amounts to this:

1. The principle explanation on offer is compatible with a number of different constructive interpretations of QM and will not be nullified or made redundant by any of them, just as with SR and thermodynamics. Thus, the principle explanation on offer is fundamental in the sense that it is more general, universal and autonomous than any particular constructive explanation or interpretation.
2. As with the case of SR, the principle explanation on offer suggests the possibility that there will never be and need never be, any constructive theory to underwrite it or subsume it. Dynamical and causal bias aside, there is no reason to rule out this possibility a priori, and SR looks to be such a case already. Thus, the principle explanation herein would be fundamental in the sense that it does not even in principle reduce to some constructive theory or explanation.

One might ask, does our principle explanation at least rule out any particular constructive interpretations of QM, or make them redundant? To which we would reply, does thermodynamics rule out or make redundant statistical mechanics or particular alternative microphysical theories? Is the converse true? Does SR rule out or make redundant alternative constructive accounts about phenomena such as Lorentz contractions? Is the converse true? Regardless of one's larger metaphysical commitments, the consensus answer to all these questions is in the negative. Let us return the focus to QM. The Lagrangian and Hamiltonian formulations of QM do not rule each other out or make one another redundant. And this claim is perfectly compatible with a psi-epistemic account of the wavefunction (see the Postscript for more details). What a principle account does is constrain constructive theories, beyond the constraints in question, it does not necessarily rule them out or make them redundant. However, a principle theory can make constructive accounts redundant, as with the case of SR and the luminiferous ether. But as we note in the Discussion, while people have abandoned theories of the luminiferous ether, some still insist there must be an underlying constructive explanation for relativistic effects such as length contraction. We disagree, but we have no way of ruling out this possibility in principle. However, as we noted above in point 1, even if there is such a constructive explanation forthcoming, SR would still be fundamental in the sense of generality, universality and autonomy.

Regardless of where one stands on these matters, there is no denying the fact that SR, with its principle explanation, has led to profound advancements in physics and we are offering a similar possibility for QM. Of course, this is not to say that our principle account creates no tension whatsoever with certain constructive accounts of, say, EPR correlations. If one is willing to accept the possibility that a principle explanation such as ours will never be reduced or underwritten by a causal or dynamical constructive account, then we have a completely local, adynamical and acausal explanation for EPR correlations that dissolves any tension between QM and SR. In short, if we are right and our principle explanation is fundamental as characterized by point 2 above, then we simply do not need constructive non-local accounts of EPR correlations, such as Bohmian mechanics and spontaneous collapse accounts. [See the Postscript for our broader interpretative commitments regarding QM.]

However, our own view aside, for those who hold that fundamental explanation must be constructive and realist in Einstein's sense of those words, none of the mainstream interpretations neatly fit the bill. Not only do most interpretations entail some form of QM holism, contextuality, and/or non-locality, the remainder invoke priority monism and/or multiple branches or outcomes. The problem with attempting a constructive account of

QM is, as articulated by Van Camp, "Constructive interpretations are attempted, but they are not unequivocally constructive in any traditional sense" [44]. Thus, he states [44]:

> The interpretive work that must be done is less in coming up with a constructive theory and thereby explaining puzzling quantum phenomena, but more in explaining why the interpretation counts as explanatory at all given that it must sacrifice some key aspect of the traditional understanding of causal-mechanical explanation.

If statistical mechanics is the paradigm example of constructive explanation, then it is hard to imagine Einstein would approve of any mainstream interpretations of QM.

Let us also note again that contrary to certain others, we are arguing that principle explanation need not ever be discharged by a constructive explanation or interpretation–causal or otherwise in SR [45–48] or in QM [49]. For example, our principle explanation avoids the complaints about Bub's proposed principle explanation of QM leveled by Felline [49]. That is, the principle being posited herein does not require a solution to the measurement problem nor again does it necessarily beg for a constructive counterpart.

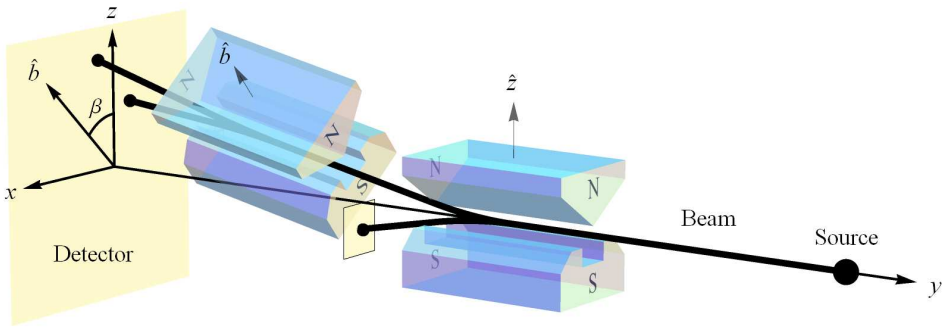


Figure 1. A pair of Stern–Gerlach (SG) spin measurements each showing the two possible outcomes, up ($+\frac{\hbar}{2}$) and down ($-\frac{\hbar}{2}$) or $+1$ and -1, for short. In this set up, the first SG magnets (oriented at $\hat{z}$) are being used to produce an initial state $|\psi\rangle = |u\rangle$ for measurement by the second SG magnets (oriented at $\hat{b}$). An important point to note here is that the classical analysis predicts all possible deflections between the target points on the detector, not just the two that are observed. The difference between the classical prediction and the quantum reality uniquely distinguishes the quantum joint distribution from the classical joint distribution for the Bell spin states [50].

To be specific, we extend NPRF from its application to the measurement of the speed of light c to include the measurement of another fundamental constant of nature, Planck's constant h. As Weinberg has noted, measuring an electron's spin via Stern–Gerlach (SG) magnets constitutes the measurement of "a universal constant of nature, Planck's constant" [51] (Figure 1). Thus, if NPRF applies equally here, everyone must measure the same value for Planck's constant h regardless of their SG magnet orientations relative to the source, which is an "empirically discovered" fact just like the light postulate. By "relative to the source," we mean relative to the plane perpendicular to the particle beam (Figure 1). In this case, the spin outcomes $\pm\frac{\hbar}{2}$ represent fundamental (indivisible) units of information per Dakic and Brukner's first axiom in their reconstruction of quantum theory, "An elementary system has the information carrying capacity of at most one bit" [43]. Therefore, the different SG magnet orientations relative to the source constitute different "reference frames" in QM, just as the different velocities relative to the source constitute different "reference frames" in SR.

To make the analogy more explicit, one could have employed NPRF to predict the light postulate as soon as Maxwell showed electromagnetic radiation propagates at $c = \frac{1}{\sqrt{\mu_o \epsilon_o}}$. All they would have had to do is extend the relativity principle from mechanics to electromagnetism. However, given the understanding of waves at the time, everyone rather

began searching for a propagation medium, i.e., the luminiferous ether. Likewise, one could have employed NPRF to predict spin angular momentum as soon as Planck published his wavelength distribution function for blackbody radiation. All they would have had to do is extend the relativity principle from mechanics and electromagnetism to blackbody radiation. However, given the understanding of angular momentum and magnetic moments at the time, Stern and Gerlach rather expected to see their silver atoms deflected in a continuum distribution after passing through their magnets (Figure 1). In other words, they discovered spin angular momentum when they were simply looking for angular momentum. However, had they noticed that their measurement constituted a measurement of Planck's constant (with its dimension of angular momentum), they could have employed NPRF to predict the spin outcome with its qubit Hilbert space structure (Figures 1 and 2) and its ineluctably probabilistic nature, as we detail in Section 3.

We can certainly imagine a world where NPRF did not apply to c and h. In the former case, c would only be measured in the "hidden" preferred frame of the luminiferous ether. In that case, the kinematic and causal structure of Minkowski spacetime would not obtain. In the latter case, h would only be measured in the "hidden" preferred frame of the orientation of the electron's angular momentum. In that case, the non-Boolean qubit Hilbert space structure would not obtain. Bub and Pitowski have pointed out the analogy between Minkowski spacetime and Hilbert space [52–54] in an attempt to explain EPR correlations. Bub sums it up nicely [55]:

> Hilbert space as a projective geometry (i.e., the subspace structure of Hilbert space) represents the structure of the space of possibilities and determines the kinematic part of quantum mechanics. ... The possibility space is a non-Boolean space in which there are built-in, structural probabilistic constraints on correlations between events (associated with the angles between the rays representing extremal events) – just as in special relativity the geometry of Minkowski space-time represents spatio-temporal constraints on events. These are kinematic, i.e., pre-dynamic, objective probabilistic or information-theoretic constraints on events to which a quantum dynamics of matter and fields conforms, through its symmetries, just as the structure of Minkowski space-time imposes spatio-temporal kinematic constraints on events to which a relativistic dynamics conforms.

But as a mere analogy, it lacks explanatory power. Herein we complete their explanatory project by showing why both aspects of their analogy follow from a common principle, NPRF.

Since QIT reconstructions of QM are based fundamentally in composite fashion on the qubit [42,43], the "very crisp and clear physical principle" of NPRF underwriting the qubit Hilbert space structure therefore underwrites the QIT reconstructions of QM. This advances QM from a mere operational theory to a proper principle theory, at least Hardy and Dakic and Brukner's reconstructions thereof. Indeed, NPRF as the physical principle behind the reconstructions of QM provides more than a mere analogy between the Lorentz transformations and the postulates of SR. That is, NPRF is to the QIT reconstructions of QM as NPRF is to the Lorentz transformations of SR. And, the fundamental transformation for the qubit at the foundation of QIT reconstructions is SO(3) [43], so we see that SO(3) and the Lorentz boosts close as a transformation group (the restricted Lorentz group) relating different reference frames in QM and SR, respectively. This also motivated Dakic and Brukner's axiom 3, which was "assumed alone for the purposes that the set of transformations builds a group structure" [43].

Essentially, we resolve the primary problem with QIT attempts to "find physical principles behind quantum theory," i.e., that they "either fail to single out the theory uniquely or are based on highly abstract mathematical assumptions without an immediate physical meaning," by explaining the qubit Hilbert space structure using constraints on QM processes in spacetime, i.e., "average-only" projection and "average-only" conservation per NPRF, rather than the converse. Thus, analogous with the structure of spacetime in SR,

our principle account of QM shows how the qubit Hilbert space structure follows from the relativity principle in spacetime, as opposed to the converse.

At the outset of Section 3, we articulate the connection between NPRF and the qubit Hilbert space structure, quantum contextuality, and the ineluctably probabilistic nature of QM. We then extrapolate this result to biparite entangled qubit systems to show why the mystery of Bell state entanglement results from conservation per NPRF in Sections 3.1 and 3.2. This will make it clear how conservation per NPRF rules out what Dakic and Brukner call "mirror quantum mechanics" in their reconstruction of QM.

3. QM from NPRF Whence Bell State Entanglement

We will refer explicitly to SG spin measurements for visualization purposes, but this can be understood to represent any measurement with a binary outcome in the symmetry plane. The only other outcome pair would be perpendicular to the symmetry plane, as in "V" $(+1)$ or "H" (-1) outcomes with photons and polarizers, in which case one thinks of "intensity of the transmitted beam" rather than "projection of the transmitted vector" [32]. The binary outcome still represents the invariant measure of the fundamental unit of action h with respect to the SO(3) transformations between QM reference frames, as in all quantum exchanges [34]. Again, SO(3) with Lorentz boosts then complete the restricted Lorentz transformation group between reference frames. As shown explicitly by Dakic and Brukner [43], the SO(3) transformation group uniquely identifies the fundamental probability structure of QM amid those of classical probability theory and higher-dimensional generalized probability theories.

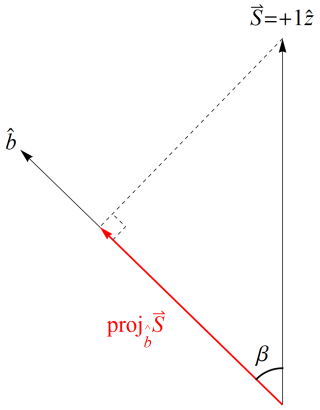


Figure 2. The spin angular momentum of Bob's particle $\vec{S}$ projected along his measurement direction $\hat{b}$. This does *not* happen with spin angular momentum due to no preferred reference frame (NPRF).

If we create a preparation state oriented along the positive z axis as in Figure 1, i.e., $|\psi\rangle = |u\rangle$, our spin angular momentum is $\vec{S} = +1\hat{z}$ (in units of $\frac{\hbar}{2} = 1$). Now proceed to make a measurement with the SG magnets oriented at $\hat{b}$ making an angle β with respect to $\hat{z}$ (Figure 1). According to classical physics, we expect to measure $\vec{S} \cdot \hat{b} = \cos(\beta)$ (Figure 2), but we cannot measure anything other than ± 1 due to NPRF (contra the prediction by classical physics), so we see that NPRF answers Wheeler's "Really Big Question," "Why the quantum? [56,57]" in "one clear, simple sentence" to convey "the central point and its necessity in the construction of the world." As a consequence, we can only recover $\cos(\beta)$ *on average* (Figure 3), i.e., NPRF dictates "average-only" projection

$$(+1)P(+1 \mid \beta) + (-1)P(-1 \mid \beta) = \cos(\beta) \quad (1)$$

Solving simultaneously with $P(+1 \mid \beta) + P(-1 \mid \beta) = 1$, we find that

$$P(+1 \mid \beta) = \cos^2\left(\frac{\beta}{2}\right) \tag{2}$$

and

$$P(-1 \mid \beta) = \sin^2\left(\frac{\beta}{2}\right) \tag{3}$$

When talking about the longitudinal outcomes [58] ("click" or "no click"), we have

$$P(V \mid \beta) = \cos^2(\beta) \tag{4}$$

and

$$P(H \mid \beta) = \sin^2(\beta) \tag{5}$$

so that our average outcome at β (orientation of polarizer with respect to initial polarization state) is given by

$$(+1)\cos^2(\beta) + (-1)\sin^2(\beta) = \cos^2(\beta) - \sin^2(\beta) \tag{6}$$

This is the naively expected Malus law per classical physics for the intensity of electromagnetic radiation transmitted through a polarizer if "pass" is $+1$ and "no pass" is -1 (instead of 0). As with the transverse mode NPRF rules out "fractional outcomes," again contra the prediction by classical physics, so the classical result obtains only on average when $\beta \neq 0$. This explains the ineluctably probabilistic nature of QM, as pointed out by Mermin [59]:

> Quantum mechanics is, after all, the first physical theory in which probability is explicitly not a way of dealing with ignorance of the precise values of existing quantities.

So, we have answered Lewis' question cited earlier, "What does the probability represent?" [30] (p. 188). Of course, these "average-only" results due to "no fractional outcomes per NPRF" hold precisely for the qubit Hilbert space structure of QM.

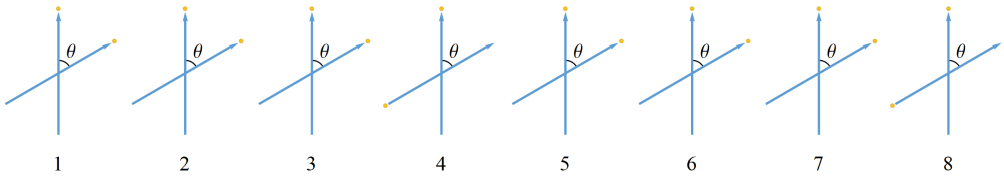


Figure 3. A spatiotemporal ensemble of 8 SG measurement trials. The blue arrows depict SG magnet orientations and the yellow dots represent the two possible measurement outcomes for each trial, up (located at arrow tip) or down (located at bottom of arrow). The vertical arrow can represent an initial state $|\psi\rangle = |u\rangle$ in which case the other arrow represents an SG measurement at $\theta = 60^\circ$ of $|\psi\rangle$. In that case, we see that the average of the ± 1 outcomes equals the projection of the initial spin angular momentum vector $\vec{S} = +1\hat{z}$ in the measurement direction $\hat{b}$, i.e., $\vec{S} \cdot \hat{b} = \cos(60^\circ) = \frac{1}{2}$. The figure can also depict two SG measurements of a spin triplet state showing Bob's(Alice's) outcomes corresponding to Alice's(Bob's) $+1$ outcomes when $\theta = 60^\circ$. For the triplet state measurements, spin angular momentum is not conserved in any given trial, because there are two different measurements being made, i.e., outcomes are in two different reference frames, but it is conserved on average for all 8 trials (six up outcomes and two down outcomes average to $\cos(60^\circ) = \frac{1}{2}$). It is impossible for spin angular momentum to be conserved explicitly in each trial since the measurement outcomes are binary (quantum) with values of $+1$ (up) or -1 (down) per NPRF. The "SO(3) conservation" at work here does not assume Alice and Bob's measured values of spin angular momentum are mere components of some hidden angular momentum (Figure 2). That is, the measured values of spin angular momentum *are* the angular momenta contributing to this "SO(3) conservation."

We ask for the reader's indulgence while we explicitly review how the qubit Hilbert space structure represented by the Pauli spin matrices evidences the relationship between

quantum contextuality and NPRF implicitly. In the eigenbasis of σ_z the Pauli spin matrices are

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -\mathbf{i} \\ \mathbf{i} & 0 \end{pmatrix}, \quad \text{and} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

where $\mathbf{i} = \sqrt{-1}$. All spin matrices have the same ± 1 eigenvalues (measurement outcomes), which reflects the fact that there are no fractional outcomes per NPRF. We denote the corresponding eigenvectors (eigenstates) as $|u\rangle$ and $|d\rangle$ for spin up $(+1)$ and spin down (-1), respectively. Using the Pauli spin matrices supra with $|u\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $|d\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, we see that $\sigma_z|u\rangle = |u\rangle$, $\sigma_z|d\rangle = -|d\rangle$, $\sigma_x|u\rangle = |d\rangle$, $\sigma_x|d\rangle = |u\rangle$, $\sigma_y|u\rangle = \mathbf{i}|d\rangle$, and $\sigma_y|d\rangle = -\mathbf{i}|u\rangle$. If we change the orientation of a vector from right pointing (ket) to left pointing (bra) or vice-versa, we transpose and take the complex conjugate. For example, if $|A\rangle = \mathbf{i}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{i}|u\rangle$, then $\langle A| = -\mathbf{i}\begin{pmatrix} 1 & 0 \end{pmatrix} = -\mathbf{i}\langle u|$. Therefore, any spin matrix can be written as $(+1)|u\rangle\langle u| + (-1)|d\rangle\langle d|$ where $|u\rangle$ and $|d\rangle$ are their up and down eigenstates, respectively. A qubit is then constructed from this two-level quantum system, i.e., $|\psi\rangle = c_1|u\rangle + c_2|d\rangle$ where $|c_1|^2 + |c_2|^2 = 1$.

An arbitrary spin measurement σ in the $\hat{b}$ direction is given by the spin matrices

$$\sigma = \hat{b} \cdot \vec{\sigma} = b_x\sigma_x + b_y\sigma_y + b_z\sigma_z \tag{7}$$

Again, preparation states $|\psi\rangle$ are created from linear combinations of the Pauli spin eigenstates. The average outcome (all we can obtain per NPRF) for a measurement σ on state $|\psi\rangle$ is given by

$$\langle\sigma\rangle := \langle\psi|\sigma|\psi\rangle \tag{8}$$

For example, in Figure 1 we have $|\psi\rangle = |u\rangle$ (prepared by the first SG magnets) and $\sigma = \sin(\beta)\sigma_x + \cos(\beta)\sigma_z$ (per the second SG magnets), so $\langle\sigma\rangle = \cos(\beta)$ in accord with Equation (1).

Finally, the probability of obtaining a $+1$ or -1 result for σ is just

$$P(+1 \mid \beta) = |\langle\psi|\tilde{u}\rangle|^2 = \cos^2\left(\frac{\beta}{2}\right) \tag{9}$$

and

$$P(-1 \mid \beta) = |\langle\psi|\tilde{d}\rangle|^2 = \sin^2\left(\frac{\beta}{2}\right) \tag{10}$$

where $|\tilde{u}\rangle$ and $|\tilde{d}\rangle$ are the eigenvectors of σ and $\frac{\beta}{2}$ is the angle between $|\psi\rangle$ and $|\tilde{u}\rangle$ in Hilbert space. This agrees with the result from NPRF in Equations (2) and (3). Thus, per Einstein's definition of a principle theory, "we have an empirically discovered principle that gives rise to mathematically formulated criteria which the separate processes or the theoretical representations of them have to satisfy."

Again, the Pauli spin matrices are created from the possible measurement outcomes ± 1 and the outer products of their eigenstates. Thus, we see that the entire qubit state and measurement structure is operationally self-referential (contextual) in that the preparation states and the measurement operators are not independent. We also see how the principle of NPRF underwrites the QM operational structure for qubits and, therefore, the QIT reconstructions of QM built upon the qubit. In the following, we will review the SU(2)/SO(3) transformation property for qubits via their bipartite entanglement in the Bell spin states.

3.1. The Bell Spin States

With that review of the implicit contextuality in the qubit operational formalism and its basis in NPRF, let us explore the conservation being depicted by the Bell spin states and relate it to the correlation function. When considering two-particle states, we will

use the juxtaposed notation for our spin states and matrices. Thus, $\sigma_x\sigma_z|ud\rangle = -|dd\rangle$ and $\sigma_x\sigma_y|ud\rangle = -\mathbf{i}|du\rangle$, for example. Essentially, we are simply ignoring the tensor product sign $\otimes$, so that $(\sigma_x \otimes \sigma_z)||u\rangle \otimes |d\rangle\rangle = \sigma_x\sigma_z|ud\rangle$. It is still easy to see which spin matrix is acting on which Hilbert space vector via the juxtaposition.

The Bell spin states are (again, omitting $\otimes$)

$$\begin{aligned} |\psi_-\rangle &= \frac{|ud\rangle - |du\rangle}{\sqrt{2}} \\ |\psi_+\rangle &= \frac{|ud\rangle + |du\rangle}{\sqrt{2}} \\ |\phi_-\rangle &= \frac{|uu\rangle - |dd\rangle}{\sqrt{2}} \\ |\phi_+\rangle &= \frac{|uu\rangle + |dd\rangle}{\sqrt{2}} \end{aligned} \tag{11}$$

in the eigenbasis of σ_z. The first state $|\psi_-\rangle$ is called the "spin singlet state" and it represents a total conserved spin angular momentum of zero ($S = 0$) for the two particles involved. The other three states are called the "spin triplet states" and they each represent a total conserved spin angular momentum of one ($S = 1$, in units of $\hbar = 1$). In all four cases, the entanglement represents the conservation of spin angular momentum for the process creating the state.

Assuming that Alice is making her spin measurement σ_1 in the $\hat{a}$ direction and Bob is making his spin measurement σ_2 in the $\hat{b}$ direction (Figure 4), we have

$$\begin{aligned} \sigma_1 &= \hat{a} \cdot \vec{\sigma} = a_x\sigma_x + a_y\sigma_y + a_z\sigma_z \\ \sigma_2 &= \hat{b} \cdot \vec{\sigma} = b_x\sigma_x + b_y\sigma_y + b_z\sigma_z \end{aligned} \tag{12}$$

Per the formalism explicated above, the correlation functions are given by (again, omitting $\otimes$)

$$\begin{aligned} \langle\psi_-|\sigma_1\sigma_2|\psi_-\rangle &= -a_xb_x - a_yb_y - a_zb_z \\ \langle\psi_+|\sigma_1\sigma_2|\psi_+\rangle &= a_xb_x + a_yb_y - a_zb_z \\ \langle\phi_-|\sigma_1\sigma_2|\phi_-\rangle &= -a_xb_x + a_yb_y + a_zb_z \\ \langle\phi_+|\sigma_1\sigma_2|\phi_+\rangle &= a_xb_x - a_yb_y + a_zb_z \end{aligned} \tag{13}$$

We now review the conservation being depicted by the Bell spin states, starting with the singlet state $|\psi_-\rangle$.

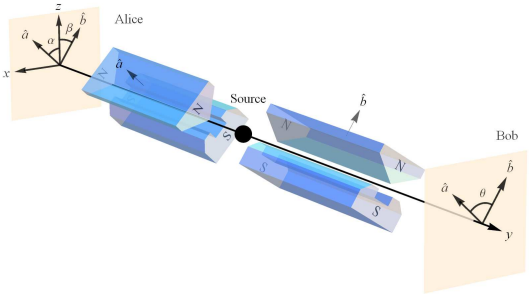


Figure 4. Alice and Bob making spin measurements on a pair of spin-entangled particles with their Stern–Gerlach (SG) magnets and detectors.

The spin singlet state is invariant under all three SU(2) transformations. For example, $|\psi_-\rangle \rightarrow |\psi_-\rangle$ when we transform our basis per

$$\begin{aligned} |u\rangle \rightarrow & \quad \cos(\Theta)|u\rangle + \sin(\Theta)|d\rangle \\ |d\rangle \rightarrow & \quad -\sin(\Theta)|u\rangle + \cos(\Theta)|d\rangle \end{aligned} \tag{14}$$

where Θ is an angle in Hilbert space (as opposed to the SG magnet angles in real space). To see what this means in real space, we construct the corresponding spin measurement operator using these transformed up $|\tilde{u}\rangle$ and down $|\tilde{d}\rangle$ vectors

$$|\tilde{u}\rangle\langle\tilde{u}| - |\tilde{d}\rangle\langle\tilde{d}| = \begin{pmatrix} \cos(2\Theta) & \sin(2\Theta) \\ \sin(2\Theta) & -\cos(2\Theta) \end{pmatrix} = \cos(2\Theta)\sigma_z + \sin(2\Theta)\sigma_x \tag{15}$$

Thus, the invariance of the state under this Hilbert space SU(2) transformation means we have rotational (SO(3)) invariance for the SG measurement outcomes in the xz-plane of real space. Specifically, $|\psi_-\rangle$ tells us that when the SG magnets are aligned in the z direction (Alice and Bob are in the same reference frame) the outcomes are always opposite ($\frac{1}{2}$ of the time ud and $\frac{1}{2}$ of the time du). Since $|\psi_-\rangle$ has that same functional form under an SU(2) transformation in Hilbert space representing an SO(3) rotation in the xz-plane per Equations (14) and (15), the outcomes are always opposite ($\frac{1}{2}$ ud and $\frac{1}{2}$ du) for aligned SG magnets in the xz-plane. That is the "SO(3) conservation" associated with this SU(2) symmetry.

Equation (15) shows us that when the angle in Hilbert space is Θ, the angle θ of the rotated SG magnets in the xz-plane is $\theta = 2\Theta$. The physical reason for this factor of 2 between Θ in Hilbert space and θ in real space can be seen in Figures 5 and 6.

Figure 5. Average View for the Spin Singlet State. Reading from left to right, as Bob rotates his SG magnets relative to Alice's SG magnets for her $+1$ outcome, the average value of his outcome varies from -1 (totally down, arrow bottom) to 0 to $+1$ (totally up, arrow tip). This obtains per conservation of spin angular momentum on average in accord with NPRF. Bob can say exactly the same about Alice's outcomes as she rotates her SG magnets relative to his SG magnets for his $+1$ outcome. That is, their outcomes can only satisfy conservation of spin angular momentum *on average* in different reference frames, because they only measure ± 1, never a fractional result. Thus, just as NPRF in SR leads to a principle explanation of time dilation and Lorentz contraction, we see that NPRF in quantum mechanics (QM) requires quantum outcomes $\pm 1\left(\frac{\hbar}{2}\right)$ for all measurements leading to a principle explanation of Bell state entanglement.

Figure 6. Average View for the Spin Triplet States. Reading from the left, as Bob(Alice) rotates his(her) SG magnets relative to Alice's(Bob's) SG magnets for her(his) $+1$ outcome, the average value of his(her) outcome varies from $+1$ (totally up, arrow tip) to 0 to -1 (totally down, arrow bottom).

Equation (14) is the Hilbert space SU(2) transformation that represents an SO(3) rotation about the y axis in real space and can be written

$$\begin{pmatrix} u \\ d \end{pmatrix} \rightarrow \begin{pmatrix} \cos(\Theta) & \sin(\Theta) \\ -\sin(\Theta) & \cos(\Theta) \end{pmatrix} \begin{pmatrix} u \\ d \end{pmatrix} = (\cos(\Theta) I + \mathbf{i} \sin(\Theta) \sigma_y) \begin{pmatrix} u \\ d \end{pmatrix} \tag{16}$$

The SU(2) transformation that represents an SO(3) rotation about the x axis in real space can be written

$$\begin{pmatrix} u \\ d \end{pmatrix} \rightarrow \begin{pmatrix} \cos(\Theta) & \mathbf{i} \sin(\Theta) \\ \mathbf{i} \sin(\Theta) & \cos(\Theta) \end{pmatrix} \begin{pmatrix} u \\ d \end{pmatrix} = (\cos(\Theta) I + \mathbf{i} \sin(\Theta) \sigma_x) \begin{pmatrix} u \\ d \end{pmatrix} \tag{17}$$

In addition, the SU(2) transformation that represents an SO(3) rotation about the z axis in real space can be written

$$\begin{pmatrix} u \\ d \end{pmatrix} \rightarrow \begin{pmatrix} \cos(\Theta) + \mathbf{i} \sin(\Theta) & 0 \\ 0 & \cos(\Theta) - \mathbf{i} \sin(\Theta) \end{pmatrix} \begin{pmatrix} u \\ d \end{pmatrix} = (\cos(\Theta) I + \mathbf{i} \sin(\Theta) \sigma_z) \begin{pmatrix} u \\ d \end{pmatrix} \tag{18}$$

The invariance of $|\psi_-\rangle$ under all three SU(2) transformations is reasonable, since the spin singlet state represents the conservation of a total, directionless spin angular momentum of $S = 0$ and each SU(2) transformation in Hilbert space corresponds to an element of SO(3) in real space. This explains why its correlation function is $-\hat{a} \cdot \hat{b}$, as shown in Equation (13). Now let us look at the spin triplet states.

Starting with $|\phi_+\rangle$, the only SU(2) transformation that takes $|\phi_+\rangle \rightarrow |\phi_+\rangle$ is Equation (14). That means this state reflects rotational (SO(3)) invariance for our SG measurement outcomes in the xz-plane. Specifically, $|\phi_+\rangle$ means when the SG magnets are aligned in the z direction (measurements are being made in the same reference frame) the outcomes are always the same ($\frac{1}{2}$ of the time uu and $\frac{1}{2}$ of the time dd). Since $|\phi_+\rangle$ has that same functional form under an SU(2) transformation in Hilbert space representing an SO(3) rotation in the xz-plane per Equations (14) and (15), the outcomes are always the same ($\frac{1}{2}$ uu and $\frac{1}{2}$ dd) for aligned SG magnets in the xz-plane. That is the "SO(3) conservation" associated with this SU(2) symmetry and it applies only for measurements made at the same angle (in the same reference frame). Here, $|\phi_+\rangle$ is only invariant under Equation (14), so we can only expect rotational invariance for our SG measurement outcomes in the xz-plane. This agrees with Equation (13) where we see that the correlation function for arbitrarily oriented σ_1 and σ_2 is $a_x b_x - a_y b_y + a_z b_z$. Therefore, unless we restrict our measurements to the xz-plane, we do not have the rotationally invariant correlation function $\hat{a} \cdot \hat{b}$ as with to the spin singlet state.

For the state $|\phi_-\rangle$, we find that the only SU(2) transformation leaving it invariant is Equation (17). Therefore, this state means we have rotational (SO(3)) invariance for the SG measurement outcomes in the yz-plane. Given that $|\phi_-\rangle$ is only invariant under Equation (17), we can only expect rotational invariance for our SG measurement outcomes in the yz-plane. This agrees with Equation (13) where we see that the correlation function for arbitrarily oriented σ_1 and σ_2 for $|\phi_-\rangle$ is given by $-a_x b_x + a_y b_y + a_z b_z$. So, unless we restrict our measurements to the yz-plane, we do not have the rotationally invariant correlation function $\hat{a} \cdot \hat{b}$ as with the spin singlet state.

Finally, $|\psi_+\rangle$ is only invariant under the SU(2) tranformation of Equation (18). Therefore, this state means we have rotational (SO(3)) invariance for our SG measurement outcomes in the xy-plane. However, unlike the situation with $|\psi_-\rangle$, we need to transform $|\psi_+\rangle$ to either the σ_x or σ_y eigenbasis to find the rotationally invariant outcome in the xy-plane. Doing so we find that the outcomes are always the same ($\frac{1}{2}$ of the time uu and $\frac{1}{2}$ of the time dd) in the xy-plane [33]. This agrees with Equation (13) where we see that the correlation function for arbitrarily oriented σ_1 and σ_2 for $|\psi_+\rangle$ is given by $a_x b_x + a_y b_y - a_z b_z$. Therefore, unless we restrict our measurements to the xy-plane, we do not have the rotationally invariant correlation function $\hat{a} \cdot \hat{b}$ as with the spin singlet state.

What does all this mean? Obviously, the SU(2) invariance of each of the spin triplet states in Hilbert space represents the SO(3) invariant conservation of spin angular momentum $S = 1$ for each of the planes xz ($|\phi_+\rangle$), yz ($|\phi_-\rangle$), and xy ($|\psi_+\rangle$) in real space. Specifically, when the measurements are being made in the same reference frame (SG magnets are aligned) anywhere in the respective symmetry plane the outcomes are always the same ($\frac{1}{2}$ of the time uu and $\frac{1}{2}$ of the time dd). That is, we have a planar conservation and our experiment would determine the plane. If you want to model a conserved $S = 1$ for some other plane, you simply expand in the spin triplet basis.

With this understanding of the conservation principle at work for entangled qubits, we see why the so-called "mirror quantum mechanics" of Dakic and Brukner [43] does not make sense physically. The "mirror" solution of their reconstruction is regular, but cannot be consistently constructed for systems of three bits. Thus, Dakic and Brukner rule it out for mathematical reasons. We can rule it out already at the level of the two-qubit system because its correlation functions are simply -1 times those in Equation (13). That means the mirror singlet state has total spin angular momentum of 1 instead of zero while the mirror triplet states have total spin angular momentum of zero instead of 1. Thus, the entire structure of rotational invariance shown above for standard QM, which makes sense physically, because $S = 0$ is directionless while $S = 1$ is directional, becomes nonsense physically in "mirror quantum mechanics."

In conclusion, we point out that the conservation at work here deals with the measurement outcomes proper. Per Dakic and Brukner's axiomatic reconstruction of quantum theory [43], the Bell spin states represent measurement outcomes on an entangled pair of "elementary systems," and "An elementary system has the information carrying capacity of at most one bit." Thus, the measurement outcomes do not represent the observed part of some hidden information carried by an underlying quantum system. Colloquially put, Alice and Bob's measurement outcomes constitute all of the available information.

3.2. NPRF and the Bell State Correlation Function

We now extrapolate our understanding of the qubit Hilbert space structure that follows from NPRF to the correlation functions for the Bell spin states of entangled qubit pairs. Assuming only that Alice and Bob each measure $+1$ and -1 with equal frequency at any arbitrary settings α and β, respectively (NPRF), the correlation function is [32,33]

$$\langle \alpha, \beta \rangle = \frac{1}{2}(+1)_A \overline{BA+} + \frac{1}{2}(-1)_A \overline{BA-} \tag{19}$$

where $\overline{BA+}$ is the average of Bob's outcomes when Alice measured $+1$ (denoted $(+1)_A$) and $\overline{BA-}$ is the average of Bob's outcomes when Alice measured -1 (denoted $(-1)_A$). That is, we have partitioned the data per Alice's equivalence relation, i.e., Alice's $+1$ results and Alice's -1 results. Note that this correlation function is independent of the formalism of QM, all we have assumed is that Alice and Bob each measure $+1$ and -1 with equal frequency for all measurement settings per NPRF. We now analyze the situation from Alice's perspective.

We will explain the case of the spin triplet state, as the case of the spin singlet state is analogous [33] (Figures 5 and 6). As with the single-particle state, classical intuition leads us to expect the projection of the spin angular momentum vector of Alice's particle $\vec{S}_A = +1\hat{a}$ along $\hat{b}$ is $\vec{S}_A \cdot \hat{b} = +\cos(\theta)$ where again θ is the angle between the unit vectors $\hat{a}$ and $\hat{b}$ (Figure 4). Again, this is because the prediction from classical physics is that all values between $+1\left(\frac{\hbar}{2}\right)$ and $-1\left(\frac{\hbar}{2}\right)$ are possible outcomes for a measurement of angular momentum. According to Alice, had Bob measured at her angle, i.e., oriented his SG magnets in the same direction, he would have found the spin angular momentum vector of his particle was $\vec{S}_B = \vec{S}_A = +1\hat{a}$ per conservation of spin angular momentum. Since he did not measure the spin angular momentum of his particle in her reference frame (same angle), he should have obtained a projected fraction of the length of $\vec{S}_B$, i.e., $\vec{S}_B \cdot \hat{b} = +1\hat{a} \cdot \hat{b} = \cos(\theta)$ (Figure 2). But according to NPRF, Bob only ever obtains $+1$

or -1 just like Alice, so he cannot measure the required fractional outcome to explicitly conserve spin angular momentum per Alice. Therefore, as with the single-particle case, NPRF means that Bob's outcomes must satisfy "average-only" projection (Figures 3 and 6), which means

$$\overline{BA+} = \cos(\theta) \tag{20}$$

Given this constraint per NPRF, as with the single-particle case, we can now use NPRF to find the joint probabilities for Alice and Bob's outcome pairs. Looking at Table 1, the rows and columns all sum to $\frac{1}{2}$ because both Alice and Bob must observe $+1$ half of the time and -1 half of the time per NPRF, which also asserts that the table is symmetric so that $P(-1,+1 \mid \theta) = P(+1,-1 \mid \theta)$. The average of Bob's outcomes given that Alice observes a $+1$ is

$$\overline{BA+} = 2P(+1,+1 \mid \theta)(+1) + 2P(+1,-1 \mid \theta)(-1) = \cos(\theta) \tag{21}$$

using conservation per NPRF. Together with the constraints on the rows/columns

$$P(+1,+1 \mid \theta) + P(+1,-1 \mid \theta) = \frac{1}{2}$$
$$P(+1,-1 \mid \theta) + P(-1,-1 \mid \theta) = \frac{1}{2},$$

we can uniquely solve for the joint probabilities

$$P(+1,+1 \mid \theta) = P(-1,-1 \mid \theta) = \frac{1}{2}\cos^2\left(\frac{\theta}{2}\right) \tag{22}$$

and

$$P(+1,-1 \mid \theta) = P(-1,+1 \mid \theta) = \frac{1}{2}\sin^2\left(\frac{\theta}{2}\right). \tag{23}$$

Now we can use these to compute $\overline{BA-}$

$$\overline{BA-} = 2P(-1,+1 \mid \theta)(+1) + 2P(-1,-1 \mid \theta)(-1) = -\cos(\theta) \tag{24}$$

Using Equations (21) and (24) in Equation (19) we obtain

$$\langle \alpha, \beta \rangle = \frac{1}{2}(+1)_A(\cos(\theta)) + \frac{1}{2}(-1)_A(-\cos(\theta)) = \cos(\theta) \tag{25}$$

which is precisely the correlation function for a spin triplet state in its symmetry plane found in Section 3.1.

Of course, Bob could partition the data according to his equivalence relation, i.e., his reference frame, so that it is Alice who must average her results, as obtained in her reference frame, to conserve spin angular momentum. Thus, the mathematical structure is again consistent with NPRF. In addition, this symmetry in perspectives requiring that Alice and Bob measure ± 1 with equal frequency for all settings, plus the average-only nature of the correlations, is precisely what precludes signalling, regardless of whether Alice's measurement settings and outcomes are spacelike or timelike related to Bob's.

Table 1. Joint probabilities for Alice and Bob's outcome pairs for the entangled particle experiment in Figure 4. The table is symmetric due to NPRF.

		Bob		
		$+1$	-1	Total
Alice	$+1$	$P(+1,+1 \mid \theta)$	$P(+1,-1 \mid \theta)$	1/2
	-1	$P(+1,-1 \mid \theta)$	$P(-1,-1 \mid \theta)$	1/2
	Total	1/2	1/2	1

Finally, since it is precisely this correlation function that is responsible for the Tsirelson bound [60–62], we see that NPRF is ultimately responsible for the Tsirelson bound. This answers Bub's question, "why is the world quantum and not classical, and why is it quantum rather than superquantum, i.e., why the Tsirelson bound for quantum correlations?" [53,63,64] (Figure 7). This also tells us why higher-dimensional generalized probability theories are not realized in Nature, i.e., the conservation principle for the fundamental two-bit system must be a qubit to accord with NPRF.

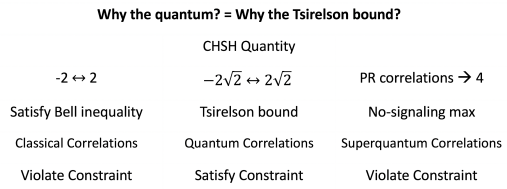


Figure 7. The "constraint" is conservation per no preferred reference frame.

4. Discussion

We have offered a principle account of EPR correlations (quantum entanglement) and quantum contextuality by applying a generalization of the relativity principle ("no preferred reference frame," NPRF) to the measurement of Planck's constant h to underwrite the qubit Hilbert space structure with its SU(2)/SO(3) transformation properties. That is, the qubit structure is the foundation of "Hilbert space as a projective geometry (i.e., the subspace structure of Hilbert space)" whence the EPR correlations. In doing so, we see that NPRF is to Hardy and Dakic and Brukner's reconstructions of QM, as NPRF is to the Lorentz transformations of SR, since the postulates of SR can be stated as NPRF applied to the measurement of the speed of light c. This answers Allori's question cited earlier, "Should I understand SSM's view as a principle theory? (But if so, which are the principles?)" [31] (p. 344).

Conservation per NPRF then accounts for no-signalling and the violations of the Bell inequality precisely to the Tsirelson bound [32], which explains why so-called "superquantum correlations" [65] and higher-dimensional generalized probability theories are not realized in Nature. Conservation per NPRF also shows so-called "mirror quantum mechanics" to be nonphysical already at the level of the two-qubit system. Thus, besides revealing a deep unity between SR and QM (Table 2), NPRF resolves many quantum mysteries.

This certainly is not what QIT had in mind for their reconstruction project. That is, they intended their reconstructions of QM would be to the "standard axioms of [quantum theory]" as "Einstein's postulates of SR are to the Lorentz transformations." As things stand now, there is no obvious connection between the interpretation-project of QM and the QIT-project [66]. In this regard, keep in mind that the postulates of SR are about the physical world in spacetime; thus, in keeping with this analogy, QIT must eventually make such correspondence to reach their lofty goals and escape the clever, but inherent instrumentalism of standard QM. Our principle account of QM, whereby NPRF is to the QIT reconstructions of QM as NPRF is to the Lorentz transformations of SR, precisely addresses the need for QIT to make correspondence with phenomena in spacetime. While these QM reconstructions do not account for all of quantum phenomena, they certainly cover bipartite qubit entanglement whence the mysteries of quantum entanglement and quantum contextuality. But, most importantly, QM reconstructions built upon the qubit Hilbert space structure explicate the essential mathematical framework for rendering QM a principle theory via NPRF.

The general idea here is that in order to make progress in the foundations of QM and in unifying QM and SR, we cannot merely continue to provide constructive empirically equivalent interpretations that lead neither to new predictions, new unifying insights, nor to underwriting QM itself. This is what we are attempting to do here. It may seem a bit

counterintuitive that NPRF underwrites both SR and QM, since quantum entanglement has been alleged by some to imply faster-than-light influences contra SR [67–69]. Per Popescu and Rohrlich [65]:

> Quantum mechanics, which does not allow us to transmit signals faster than light, preserves relativistic causality. But quantum mechanics does not always allow us to consider distant systems as separate, as Einstein assumed. The failure of Einstein separability violates, not the letter, but the spirit of special relativity, and left many physicists (including Bell) deeply unsettled.

Obviously QM (non-relativistic quantum mechanics) is not Lorentz invariant, so it certainly differs from SR in that regard. QM follows from Lorentz invariant quantum field theory only in the low energy approximation [70] (p. 173). However, claiming that SR and QM are somehow at odds based on quantum entanglement has empirical consequences, because we have experimental evidence verifying the violation of the Bell inequality in accord with quantum entanglement. Thus, if the violation of the Bell inequality is problematic for SR, then SR is being empirically challenged in some sense, hence Bell's unease.

For example, Newtonian mechanics deviates from SR because it is not Lorentz invariant. Accordingly, Newtonian mechanics predicts that velocities add in a different fashion than in SR, so imagine we found experimentally that velocities add according to Newtonian mechanics. That would not only mean Newtonian mechanics and SR are at odds, that would mean SR has been empirically refuted. Bell's unease aside, clearly, few people believe that QM has literally falsified SR. But we have gone further to show that not only is there no tension between QM and SR in substance or spirit, NPRF provides a completely local principle account of EPR correlations. Indeed, even the no-signalling feature of entangled qubits follows necessarily from NPRF. Thus, far from being incompatible, SR and QM share a deep coherence via NPRF (Table 2). This principle explanation for EPR correlations requires no violation of the causal structure of SR and it does not require the addition of a preferred frame as some non-local interpretations do, such as Bohmian mechanics and spontaneous collapse interpretations. Furthermore, this principle explanation for EPR correlations requires no causal or constructive explanation whatsoever, and that includes retrocausal mechanisms and processes. Indeed, this principle account of QM does not even require a metaphysical commitment to the block universe.

Table 2. Comparing special relativity with quantum mechanics according to no preferred reference frame (NPRF). Because Alice and Bob both measure the same speed of light c, regardless of their motion relative to the source per NPRF, Alice (Bob) may claim that Bob's (Alice's) length and time measurements are erroneous and need to be corrected (Lorentz contraction and time dilation). Likewise, because Alice and Bob both measure the same values for spin angular momentum $\pm 1\left(\frac{\hbar}{2}\right)$, regardless of their SG magnet orientation relative to the source per NPRF, Alice (Bob) may claim that Bob's (Alice's) individual ± 1 values are erroneous and need to be corrected (averaged, Figures 3, 5 and 6). In both cases, NPRF resolves the mystery it creates. In SR, the apparently inconsistent results can be reconciled via the relativity of simultaneity. That is, Alice and Bob each partition spacetime per their own equivalence relations (per their own reference frames), so that equivalence classes are their own surfaces of simultaneity and these partitions are equally valid per NPRF. This is completely analogous to QM, where the apparently inconsistent results per the Bell spin states arising because of NPRF can be reconciled by NPRF via the "relativity of data partition." That is, Alice and Bob each partition the data per their own equivalence relations (per their own reference frames), so that equivalence classes are their own $+1$ and -1 data events and these partitions are equally valid per NPRF.

Special Relativity	**Quantum Mechanics**
Empirical Fact: Alice and Bob both measure c, regardless of their motion relative to the source	Empirical Fact: Alice and Bob both measure $\pm 1\left(\frac{\hbar}{2}\right)$, regardless of their SG orientation relative to the source
Alice(Bob) says of Bob(Alice): Must correct his(her) length and time measurements	Alice(Bob) says of Bob(Alice): Must average his(her) ± 1 outcomes for projection/conservation
NPRF: Relativity of simultaneity	NPRF: Relativity of data partition

Despite the fact that this principle explanation supplies a unifying framework for both QM and SR, some might demand a constructive explanation with its corresponding "knowledge of how things in the world work, that is, of the mechanisms (often hidden) that produce the phenomena we want to understand" [71] (p. 15). This is "the causal/mechanical view of scientific explanation" per Salmon [71] (p. 15). Thus, as with SR, not everyone will consider our principle account of QM to be explanatory since, "By its very nature such a theory-of-principle explanation will have nothing to say about the reality behind the phenomenon" [72] (p. 331). As Lorentz famously complained about SR [73] (p. 230):

> Einstein simply postulates what we have deduced, with some difficulty and not altogether satisfactorily, from the fundamental equations of the electromagnetic field.

And, Albert Michelson said [74]:

> It must be admitted, these experiments are not sufficient to justify the hypothesis of an ether. But then, how can the negative result be explained?

In other words, neither was convinced that NPRF was sufficient to explain time dilation and Lorentz contraction. More recently Brown has made a similar claim [46] (p. 76):

> What has been shown is that rods and clocks must behave in quite particular ways in order for the two postulates to be true together. But this hardly amounts to an explanation of such behaviour. Rather things go the other way around. It is because rods and clocks behave as they do, in a way that is consistent with the relativity principle, that light is measured to have the same speed in each inertial frame.

In other words, the assumption is that the true or fundamental "explanation" of EPR correlations must be a constructive one in the sense of adverting to causal processes or causal mechanisms. Apparently for people with such a Reichenbachian or constructive mind-set, any principle explanation must be accounted for by some such story, e.g., the luminiferous ether. Indeed, contrary to all accepted physics, Brown and Pooley [46] have recently called for such a constructive explanation even in SR. Brown and Pooley like to make this a debate about constructive versus "geometric" explanation. They believe that the principle explanation of Lorentz contractions in SR is underwritten only by the geometry of Minkowski spacetime.

We think this misses the point, as one could believe that SR provides a principle explanation of Lorentz contractions without being a realist or a substantivialist about Minkowski spacetime. Notice, there is nothing inherently geometric about our principle explanation of EPR correlations in particular, or of NPRF in general. We would say that Brown and Pooley got it exactly wrong. It is QM that needs to become explicitly more like SR, not the other way around. Indeed, as noted in Section 2, neither textbook QM nor any of its constructive interpretations, ever made for very convincing constructive theories anyway. If QM had struck people as being like statistical mechanics, there would be no cottage industry of cooking up constructive interpretations and no need for anything like QIT reconstructions. We hope to have shed some light on why QM actually works as it does.

After Einstein published SR in 1905, physicists gradually lost interest in theories of the luminiferous ether, preferred reference frames, or any other causal account of Lorentz contractions and time dilation. Even Lorentz seemed to acknowledge the value of this principle explanation when he wrote [73] (p. 230):

> By doing so, [Einstein] may certainly take credit for making us see in the negative result of experiments like those of Michelson, Rayleigh, and Brace, not a fortuitous compensation of opposing effects but the manifestation of a general and fundamental principle.

Thus, 85 years after the publication of the EPR paper without a consensus constructive or causal account of EPR correlations, perhaps it is time to consider the possibility that

physicists will eventually likewise stop looking for constructive accounts of EPR correlations. After all, we now know that the widely accepted relativity principle is precisely the principle that resolves a plethora of QM mysteries. And, as Pauli once stated [75] (p. 33):

> 'Understanding' nature surely means taking a close look at its connections, being certain of its inner workings. Such knowledge cannot be gained by understanding an isolated phenomenon or a single group of phenomena, even if one discovers some order in them. It comes from the recognition that a wealth of experiential *facts are interconnected and can therefore be reduced to a common principle*. In that case, certainty rests precisely on this wealth of facts. The danger of making mistakes is the smaller, the richer and more complex the phenomena are, and the simpler is the common principle to which they can all be brought back. ... 'Understanding' probably means nothing more than having whatever ideas and concepts are needed to recognize that a great many different phenomena are part of a coherent whole. [Italics ours.]

One could hardly ask for a "simple common principle" or "one clear, simple sentence" more compelling than "no preferred reference frame" to convey "the central point and its necessity in the construction of the world." Perhaps causal accounts of quantum entanglement and quantum contextuality are destined to share the same fate as theories of the luminiferous ether. Perhaps this principle account will finally cause us to let go of the Reichenbachian past, and go back to the future with Einstein's insights about principle explanation.

Postscript

We understand how deep the causal and dynamical explanatory bias goes and thus we know that people will feel like we've dodged something important by not saying more about our ontology/beables and more about what the role of the wavefunction is on our account. That is, our answer to the challenge to causality by the EPR correlations resides in the probability structure of QM alone, "the kinematic part of QM" per Bub [55], so we have not offered anything concerning QM dynamics. And of course when people demand to know your "ontology," they mean that in the constructive, causal and dynamical sense of the word. Part of the causal and dynamical bias, the very reason many people think constructive explanation must be fundamental, is because they assume the world is composed of or otherwise determined by such beables. That is, they want to know what the world is made of or what matter is. Herein, we will briefly provide our answers to those questions by explaining how our principle account of QM is a realist, psi-epistemic account. Let us note however that one need not share all our commitments herein to find our preceding principle account compelling.

Let us begin with the psi-epistemic part. Explaining what this means for us requires going beyond the scope of this paper, as the question of Schrödinger dynamics differs from the question of the probability structure of QM [8,22]. If one constructs the differential equation (Schrödinger equation) corresponding to the Feynman path integral, the time-dependent foliation of spacetime gives the wavefunction $\psi(x,t)$ in concert with our time-evolved perceptions and the fact that we do not know when the outcome is going to occur. Once one has an outcome, both the configuration x_o, that is the specific spatial locations of the experimental outcomes, and time t_o of the outcomes are fixed, so the wavefunction $\psi(x,t)$ of configuration space becomes a probability amplitude $\psi(x_o,t_o)$ in spacetime, i.e., a probability amplitude for a specific outcome in spacetime. Again, the evolution of the wavefunction in configuration space before it becomes a probability amplitude in spacetime is governed by the Schrödinger equation.

However, the abrupt change from wavefunction in configuration space to probability amplitude in spacetime is not governed by the Schrödinger equation. In fact, if the Schrödinger equation is universally valid, it would simply say that the process of measurement should entangle the measurement device with the particle being measured, leaving them both to evolve according to the Schrödinger equation in a more complex configuration

space (as in the relative-states formalism shown below). Certain interpretations of QM notwithstanding, we do not seem to experience such entangled existence in configuration space, which would contain all possible experimental outcomes. Instead, we experience a single experimental outcome in spacetime.

This contradiction between theory and experience is called the "measurement problem." However, the time-evolved story in configuration space is not an issue with the path integral formalism as we interpret it, because we compute $\psi(x_o, t_o)$ directly. That is, in asking about a specific outcome we must specify the future boundary conditions that already contain definite and unique outcomes. Thus, the measurement problem is a non-starter for us. When a QM interpretation assumes the wavefunction is an epistemological tool rather than an ontological entity, that interpretation is called "psi-epistemic." In our path integral, contextuality-based account the wavefunction in configuration space is not even used, so our account is trivially psi-epistemic. Thus, our account of the wavefunction is very much like Rovelli's as we described it earlier in the paper. Our view is a complete rejection of Hilbert space realism and the like. In short, we would say that the operational recipe of textbook Hamiltonian-based QM with its Schrödinger dynamics, is from the ant's-eye perspective, the very best one could do given that the primary goal is prediction of the temporal evolution of the QM state, regardless of its ultimate ontic status.

This covers the psi-epistemic part of our view, but what about the realist part? Of course, again, this begs the question of what is required for a realist account of QM. Let us begin with the obvious. Unless one is begging the question, there is nothing inherently anti-realist or subjectivist about principle explanation. After all, NPRF generally, and "average-only" conservation specifically, are real, mind-independent and perspective-independent facts about spacetime. Indeed, by this measure, our principle explanation for EPR correlations and the like, is much more realist than the retrocausal perspectival causal explanations and also more realist than most of the psi-epistemic accounts on offer. This bears repeating. What is doing the explanatory work on our principle account of QM, just as with SR, are mind and perspective-independent facts about the world.

Again, one might nonetheless feel that our account fails to be realist because it does not provide a specific constructive ontology, such as particles, fields or waves. The first thing to note here is that the whole point of principle explanation is that it is compatible with any number of such ontologies, and that it is not incumbent upon the purveyor of such explanations to provide a constructive ontology, because such an ontology is not relevant to the explanation at hand. Indeed, as we have pointed out on numerous occasions, our principle explanation of EPR correlations is even compatible with unmediated exchanges/direct action between source and detector, i.e., the idea that there are no world-lines of counterfactual definiteness that connect the source and the detector. But we get it, many people will feel that something is missing if we cannot say exactly, in constructive terms, what goes on between source and detector. From their perspective, so far, the only thing we have told you is that the wavefunction and Hilbert space are not real, but not, what is "real." Is it particles, fields, waves in spacetime, or something else?

This is not an easy question to answer for many reasons. First, as is well known, the standard definitions that provide the essence of "particles," "waves" and "fields" are all violated by one or another weirdness of QM, e.g., that particles are strictly point like, that fields are a fully continuous and contiguous medium with all definite values at every point in spacetime for which counterfactual definiteness obtains, and that waves (in spacetime at least) must be fully and always wave-like in their behavior and be instantiated in some material or energetic medium. Second, if contextuality is a fundamental fact about the world as seems to be the case based on experimental evidence and several theorems, as we noted earlier, this calls into question the very idea of classical objects composed of or realized by autonomous self-existent QM entities with definite, intrinsic properties, and what Einstein called "primitive thisness," or what is sometimes called haecceity.

In accord with Rovelli's relational QM, our conjecture, given all of the above, is that the search for such a fundamental context-free ontology is misguided. Indeed, both relational

QM and our view are inspired by the lessons of SR, that certain facts, entities and quantities are reference frame dependent, and we also attempt to apply that idea to QM. There are certainly other similarities as well, e.g., the idea that QM is complete, the psi-epistemic take, the focus on information, and the invocation of contextuality. However, in order to explain the EPR correlations, relational QM merely defers to the entangled qubit Hilbert space structure, while QIT note that the mystery is, "Why the qubit structure? Why not classical bits? Or, generalized higher-dimensional bits? The no-signalling requirement does not suffice to rule them out." And, that answer is not to be found in the conservation represented by the Bell spin states when the measurements are made in the same reference frame (SG magnets are at the same angle). Those perfect correlations are easily replicated by assuming "hidden, definite values." The answer resides in the conservation represented by the Bell spin states when the measurements are made in *different* reference frames (SG magnets are at different angles). It is there that one finds "average-only" conservation per NPRF due to "average-only" projection per NPRF, as explained in Section 3. By taking seriously NPRF and not just relationalism, we have underwritten QM and explained its informational structure without giving up The Absoluteness of Observed Events, as is entailed by relational QM in certain cases. NPRF and The Absoluteness of Observed Events is the very heart of SR, and the very basis for its relationalism. Our diagnosis is that Rovelli drops this insight because unlike explanation in SR, he has not fully transcended the dynamical and causal explanatory bias.

As for fundamental ontology, we would say that multiscale contextuality itself is fundamental, and sometimes, depending on various contextual features, reality (whatever its ultimate *metaphysical* nature), behaves in a particle-like, field-like or wave-like fashion. We think the twin-slit experiment alone is sufficient to see how this might be so. While all this is beyond the scope of our paper, in our view, environmental decoherence, so called QM non-separability, so called QM holism, so called QM relationalism, QM dispositionalism, etc., are really just symptomatic of the fundamentality of multiscale contextuality. And furthermore, while the contextuality in question is often manifested in dynamical and causal interactions, the deeper contextuality that explains and underwrites certain aspects of those interactions is sometimes non-causal, non-dynamical and spatiotemporal–what we call adynamical global constraints. For example, as we demonstrated herein, the kind of contextuality we see in the case of EPR correlations is a consequence of NPRF. We see nothing inherently anti-realist about this view, as again, conservation laws and multiscale contextuality are real mind and perspective-independent facts about the world. What QM and relativity are really telling us is that to exist, i.e., to be a diachronic entity in space and time, is to interact with the rest of the universe creating a consistent, shared set of classical information constituting the universe [8,34]. Again, this is another way in which our view is a fully realist one, given that NPRF is at the core of our account of the physical world [8,34], as with SR itself, The Absoluteness of Observed Events can never be violated on our view.

Author Contributions: Conceptualization, M.S., W.M.S., and T.M.; formal analysis, W.M.S., T.M., and M.S.; Writing—Original draft preparation, W.M.S.; Writing—Review and editing, M.S., W.M.S., and T.M.; supervision, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Friederich, S.; Evans, P. Retrocausality in Quantum Mechanics. 2019. Available online: https://plato.stanford.edu/archives/sum2019/entries/qm-retrocausality (accessed on 12 January 2021).

2. Stuckey, W.; Silberstein, M.; McDevitt, T. Relational Blockworld: Providing a Realist Psi-Epistemic Account of Quantum Mechanics. *Int. J. Quantum Found.* **2015**, *1*, 123–170. Available online: http://www.ijqf.org/wps/wp-content/uploads/2015/06/IJQF2015v1n3p2.pdf (accessed on 12 January 2021).
3. Healey, R. Pragmatist Quantum Realism. In *Scientific Realism And The Quantum*; French, S., Saatsi, J., Eds.; Oxford University Press: Oxford, UK, 2020; pp. 123–146.
4. Laudisa, F.; Rovelli, C. Relational Quantum Mechanics. 2019. Available online: https://plato.stanford.edu/archives/win2019/entries/qm-relational/ (accessed on 12 January 2021).
5. Bell, J. The Theory of Local Beables. 1975. Available online: https://cds.cern.ch/record/980036/files/197508125.pdf (accessed on 12 January 2021).
6. Baggott, J. *Quantum Reality: The Quest for the Real Meaning of Quantum Mechanics—A Game of Theories*; Oxford University Press: Oxford, UK, 2020.
7. Bong, K.; Utreras-Alarcón, A.; Ghafari, F.; Liang, Y.; Tischler, N.; Cavalcanti, E.; Pryde, G.; Wiseman, H. A strong no-go theorem on the Wigner's friend paradox. *Nat. Phys.* **2020**, *16*, 1199–1205. [CrossRef]
8. Silberstein, M.; Stuckey, W. The Completeness of Quantum Mechanics and the Determinateness and Consistency of Intersubjective Experience: Wigner's Friend and Delayed Choice. In *Quantum Mechanics and Consciousness*; Gao, S., Ed.; Oxford University Press: Oxford, UK, 2021.
9. Spekkens, R. Contextuality for Preparations, Transformations, and Unsharp Measurements. *Phys. Rev. A* **2005**, *71*, 052108. [CrossRef]
10. Dowe, P. *Physical Causation*; Cambridge University Press: Cambridge, UK, 2000.
11. Born, M. *The Born-Einstein letters: Correspondence between Albert Einstein and Max and Hedwig Born from 1916–1955. with Commentaries by Max Born*; Macmillan: New York, NY, USA, 1971.
12. Aharonov, Y. The Two-State Vector Formalism: An Updated Review. In *Time in Quantum Mechanics*; Muga, J., Mayato, R.S., Egusquiza, I., Eds.; Springer: Berlin, Germany, 2007; pp. 399–447.
13. Kastner, R.E. *The Transactional Interpretation of Quantum Mechanics: The Reality of Possibility*; Cambridge University Press: Cambridge, MA, USA, 2013.
14. Cramer, J. *The Quantum Handshake*; Springer: Dordrecht, The Netherlands, 2016.
15. Wharton, K. Towards a Realistic Parsing of the Feynman Path Integral. *Quanta* **2016**, *5*, 1–11. [CrossRef]
16. Wharton, K. A New Class of Retrocausal Models. *Entropy* **2018**, *20*, 410. Available online: https://arxiv.org/abs/1805.09731 (accessed on 12 January 2021). [CrossRef] [PubMed]
17. Woodward, J. *Making Things Happen: A Theory of Causal Explanation*; Oxford University Press: New York, NY, USA, 2003.
18. Pearl, J. *Causality: Models, Reasoning, and Inference*; Cambridge University Press: New York, NY, USA, 2009.
19. Price, H. Causal Perspectivalism. In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*; Price, H., Corry, R., Eds.; Oxford University Press: Oxford, UK, 2007; pp. 250–292.
20. Price, H.; Weslake, B. The Time-Asymmetry of Causation. In *The Oxford Handbook of Causation*; Beebee, H., Hitchcock, C., Menzies, P., Eds.; Oxford University Press: New York, NY, USA, 2010; pp. 414–443.
21. Evans, P. Retrocausality at No Extra Cost. *Synthese* **2015**, *192*, 1139–1155. [CrossRef]
22. Silberstein, M.; Stuckey, W.; McDevitt, T. *Beyond the Dynamical Universe: Unifying Block Universe Physics and Time as Experienced*; Oxford University Press: Oxford, UK, 2018.
23. Silberstein, M.; M Cifone, M.; Stuckey, W. Why Quantum Mechanics Favors Adynamical and Acausal Interpretations such as Relational Blockworld over Backwardly Causal and Time-Symmetric Rivals. *Stud. Hist. Philos. Mod. Phys.* **2008**, *39*, 736–751. [CrossRef]
24. Maudlin, T. *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*, 3rd ed.; Wiley-Blackwell: Malden, MA, USA, 2011.
25. Cramer, J.G. The Transactional Interpretation of Quantum Mechanics. *Rev. Mod. Phys.* **1986**, *58*, 647–687. [CrossRef]
26. Shrapnel, S.; Costa, F. Causation Does not Explain Contextuality. *Quantum* **2018**, *2*, 63. Available online: https://arxiv.org/abs/1708.00137 (accessed on 12 January 2021). [CrossRef]
27. Merali, Z. This Twist on Schrödinger's Cat Paradox Has Major Implications for Quantum Theory. 2020. Available online: https://www.scientificamerican.com/article/this-twist-on-schroedingers-cat-paradox-has-major-implications-for-quantum-theory/ (accessed on 12 January 2021).
28. Hossenfelder, S.; Palmer, T. How to Make Sense of Quantum Physics: Superdeterminism, A Long-Abandoned Idea, May Help Us Overcome the Current Crisis in Physics. 2020. Available online: http://nautil.us/issue/83/intelligence/how-to-make-sense-of-quantum-physics (accessed on 12 January 2021).
29. Price, H.; Wharton, K. Dispelling the Quantum Spooks: A Clue That Einstein Missed? In *Time of Nature and the Nature of Time*; Bouton, C., Huneman, P., Eds.; Springer: Dordrecht, The Netherlands, 2017; pp. 123–137.
30. Lewis, P. Review of "Beyond the Dynamical Universe: Unifying Block Universe Physics and Time as Experienced". *Int. J. Quantum Found.* **2019**, *5*, 186–188.
31. Allori, V. Book Review of "Beyond the Dynamical Universe: Unifying Block Universe Physics and Time as Experienced," by Michael Silberstein, W.M. Stuckey, and Timothy McDevitt. *Metascience* **2019**, *28*, 341–344. [CrossRef]

32. Stuckey, W.; Silberstein, M.; McDevitt, T.; Kohler, I. Why the Tsirelson Bound? Bub's Question and Fuchs' Desideratum. *Entropy* **2019**, *21*, 692. Available online: https://arxiv.org/abs/1807.09115 (accessed on 12 January 2021). [CrossRef] [PubMed]
33. Stuckey, W.; Silberstein, M.; McDevitt, T.; Le, T. Answering Mermin's Challenge with Conservation per No Preferred Reference Frame. *Sci. Rep.* **2020**, *10*, 15771. Available online: www.nature.com/articles/s41598-020-72817-7 (accessed on 12 January 2021). [CrossRef] [PubMed]
34. Silberstein, M.; Stuckey, W. Re-Thinking the World with Neutral Monism: Removing the Boundaries Between Mind, Matter, and Spacetime. *Entropy* **2020**, *22*, 551. [CrossRef] [PubMed]
35. Felline, L. Scientific Explanation between Principle and Constructive Theories. *Philos. Sci.* **2011**, *78*, 989–1000. [CrossRef]
36. Einstein, A. What is the Theory of Relativity? *London Times*, 28 November 1919; pp. 53–54.
37. Einstein, A. Autobiographical Notes. In *Albert Einstein: Philosopher-Scientist*; Schilpp, P.A., Ed.; Open Court: La Salle, IL, USA, 1949; pp. 3–94.
38. Mainwood, P. What Do Most People Misunderstand About Einstein's Theory Of Relativity? 2018. Available online: https://www.forbes.com/sites/quora/2018/09/19/what-do-most-people-misunderstand-about-einsteins-theory-of-relativity (accessed on 12 January 2021).
39. Einstein, A.; Podolsky, B.; Rosen, N. Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Phys. Rev.* **1935**, *47*, 777–780. [CrossRef]
40. Smolin, L. *Einstein's Unfinished Revolution: The Search for What Lies Beyond the Quantum*; Penguin Press: New York, NY, USA, 2019.
41. Fuchs, C.; Stacey, B. Some Negative Remarks on Operational Approaches to Quantum Theory. In *Quantum Theory: Informational Foundations and Foils*; Chiribella, G., Spekkens, R., Eds.; Springer: Dordrecht, The Netherlands, 2016; pp. 283–305.
42. Hardy, L. Reconstructing Quantum Theory. In *Quantum Theory: Informational Foundations and Foils*; Chiribella, G., Spekkens, R., Eds.; Springer: Dordrecht, The Netherlands, 2016; pp. 223–248. Available online: https://arxiv.org/abs/1303.1538 (accessed on 12 January 2021).
43. Dakic, B.; Brukner, C. Quantum Theory and Beyond: Is Entanglement Special? In *Deep Beauty: Understanding the Quantum World through Mathematical Innovation*; Halvorson, H., Ed.; Cambridge University Press: Cambridge, MA, USA, 2009; pp. 365–392. Available online: https://arxiv.org/abs/0911.0695 (accessed on 12 January 2021).
44. Van Camp, W. Principle Theories, Constructive Theories, and Explanation in Modern Physics. *Stud. Hist. Philos. Sci. Part Stud. Hist. Philos. Mod. Phys.* **2011**, *42*, 23–31. [CrossRef]
45. Brown, H. *Physical Relativity: Spacetime Structure from a Dynamical Perspective*; Oxford University Press: Oxford, UK, 2005.
46. Brown, H.; Pooley, O. Minkowski Space-Time: A Glorious Non-Entity. In *The Ontology of Spacetime*; Dieks, D., Ed.; Elsevier: Amsterdam, The Netherlands, 2006; p. 67.
47. Norton, J. Why Constructive Relativity Fails. *Br. J. Philos. Sci.* **2008**, *59*, 821–834. [CrossRef]
48. Menon, T. Algebraic Fields and the Dynamical Approach to Physical Geometry. *Philos. Sci.* **2019**, *86*, 1273–1283. [CrossRef]
49. Felline, L. Quantum Theory is Not Only About Information. In *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 1355–2198. Available online: https://arxiv.org/abs/1806.05323 (accessed on 12 January 2021).
50. Garg, A.; Mermin, N. Bell Inequalities with a Range of Violation that Does Not Diminish as the Spin Becomes Arbitrarily Large. *Phys. Rev. Lett.* **1982**, *49*, 901–904. [CrossRef]
51. Weinberg, S. The Trouble with Quantum Mechanics. 2017. Available online: https://www.nybooks.com/articles/2017/01/19/trouble-with-quantum-mechanics/ (accessed on 12 January 2021).
52. Bub, J.; Pitowski, I. Two dogmas about quantum mechanics. In *Many Worlds? Everett, Quantum Theory, and Reality*; Saunders, S., Barrett, J., Kent, A., Wallace, D., Eds.; Oxford University Press: Oxford, UK, 2010; pp. 431–456.
53. Bub, J. *Bananaworld: Quantum Mechanics for Primates*; Oxford University Press: Oxford, UK, 2016.
54. Bub, J. 'Two Dogmas' Redux. In *Quantum, Probability, Logic: The Work and Influence of Itamar Pitowsky*; Hemmo, M., Shenker, O., Eds.; Springer Nature: London, UK, 2020; pp. 199–215.
55. Bub, J. Quantum Correlations and the Measurement Problem. 2012. Available online: https://arxiv.org/abs/1210.6371 (accessed on 12 January 2021).
56. Wheeler, J. How Come the Quantum? *New Tech. Ideas Quantum Meas. Theory* **1986**, *480*, 304–316. [CrossRef]
57. Barrow, J.D.; Davies, P.C.W.; Charles, L.; Harper, J. (Eds.) *Science and Ultimate Reality: Quantum Theory, Cosmology, and Complexity*; Cambridge University Press: New York, NY, USA, 2004.
58. Dehlinger, D.; Mitchell, M. Entangled photons, nonlocality, and Bell inequalities in the undergraduate laboratory. *Am. J. Phys.* **2002**, *70*, 903–910. [CrossRef]
59. Mermin, N.D. Making better sense of quantum mechanics. *Rep. Prog. Phys.* **2019**, *82*, 012002. Available online: https://arxiv.org/abs/1809.01639 (accessed on 12 January 2021). [CrossRef] [PubMed]
60. Cirel'son, B. Quantum Generalizations of Bell's Inequality. *Lett. Math. Phys.* **1980**, *4*, 93–100. Available online: https://www.tau.ac.il/~tsirel/download/qbell80.pdf (accessed on 12 January 2021). [CrossRef]
61. Landau, L.J. On the violation of Bell's inequality in quantum theory. *Phys. Lett. A* **1987**, *120*, 54–56. [CrossRef]
62. Khalfin, L.A.; Tsirelson, B.S. Quantum/Classical Correspondence in the Light of Bell's Inequalities. *Found. Phys.* **1992**, *22*, 879–948. Available online: https://www.tau.ac.il/~tsirel/download/quantcl.ps (accessed on 12 January 2021). [CrossRef]
63. Bub, J. Why the Quantum? *Stud. Hist. Philos. Mod. Phys.* **2004**, *35B*, 241–266. [CrossRef]

64. Bub, J. Why the Tsirelson bound? In *The Probable and the Improbable: The Meaning and Role of Probability in Physics*; Hemmo, M., Ben-Menahem, Y., Eds.; Springer: Dordrecht, The Netherlands, 2012; pp. 167–185. Available online: https://arxiv.org/abs/1208.3744 (accessed on 12 January 2021).
65. Popescu, S.; Rohrlich, D. Quantum nonlocality as an axiom. *Found. Phys.* **1994**, *24*, 379–385. Available online: https://arxiv.org/abs/quant-ph/9508009v1 (accessed on 12 January 2021). [CrossRef]
66. Timpson, C.G. *Quantum Information Theory & the Foundations of Quantum Mechanics*; Oxford University Press: Oxford, UK, 2013.
67. Alford, M. Ghostly action at a distance: A non-technical explanation of the Bell inequality. *Am. J. Phys.* **2016**, *84*, 448–457. [CrossRef]
68. Mamone-Capria, M. On the Incompatibility of Special Relativity and Quantum Mechanics. *J. Found. Appl. Phys.* **2018**, *8*, 163–189. Available online: https://arxiv.org/pdf/1704.02587.pdf (accessed on 12 January 2021).
69. Bell, J. *Speakable and Unspeakable in Quantum Mechanics*; Cambridge University Press: Cambridge, MA, USA, 1987.
70. Zee, A. *Quantum Field Theory in a Nutshell*; Princeton University Press: Princeton, NJ, USA, 2003.
71. Salmon, W.C. The value of scientific understanding. *Philosophica* **1993**, *51*, 9–19.
72. Balashov, Y.; Janssen, M. Presentism and Relativity. *Br. J. Philos. Sci.* **2003**, *54*, 327–346. [CrossRef]
73. Lorentz, H. *The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat*; G.E. Stechert and Co.: New York, NY, USA, 1916.
74. Bane, D. The Mechanical Universe Episode 41: The Michelson-Morley Experiment, 1985. Albert Michelson quote from 1931. Available online: https://www.teacherspayteachers.com/Product/The-Mechanical-Universe-Episode-41-The-Michelson-Morley-Experiment-5122993 (accessed on 12 January 2021).
75. Heisenberg, W. *Physics and Beyond: Encounters and Conversations*; Harper & Row: New York, NY, USA, 1971.

A Principle Explanation of Bell State Entanglement

W.M. Stuckey* Michael Silberstein† ‡ and Timothy McDevitt§

2 February 2021

Abstract

Many in quantum foundations seek a principle explanation of Bell state entanglement. While reconstructions of quantum mechanics (QM) have been produced, the community does not find them compelling. Herein we offer a principle explanation for Bell state entanglement, i.e., conservation per no preferred reference frame (NPRF), such that NPRF unifies Bell state entanglement with length contraction and time dilation from special relativity (SR). What makes this a principle explanation is that it's grounded directly in phenomenology, it is an adynamical and acausal explanation that involves adynamical global constraints as opposed to dynamical laws or causal mechanisms, and it's unifying with respect to QM and SR.

1 Introduction

Many physicists in quantum information theory (QIT) are calling for "clear physical principles" [Fuchs and Stacey, 2016] to account for quantum mechanics (QM). As [Hardy, 2016] points out, "The standard axioms of [quantum theory] are rather ad hoc. Where does this structure come from?" Fuchs and Hardy point to the postulates of special relativity (SR) as an example of what QIT seeks for QM [Fuchs and Stacey, 2016, Hardy, 2016] and SR is a principle theory [Felline, 2011]. That is, the postulates of SR are constraints offered without a corresponding constructive explanation. In what follows, [Einstein, 1919] explains the difference between the two:

> We can distinguish various kinds of theories in physics. Most of them are constructive. They attempt to build up a picture of the more complex phenomena out of the materials of a relatively simple

*Department of Physics, Elizabethtown College, Elizabethtown, PA 17022, USA
†Department of Philosophy, Elizabethtown College, Elizabethtown, PA 17022, USA
‡Department of Philosophy, University of Maryland, College Park, MD 20742, USA
§Department of Mathematical Sciences, Elizabethtown College, Elizabethtown, PA 17022, USA

> formal scheme from which they start out. [The kinetic theory of gases is an example.] ...
>
> Along with this most important class of theories there exists a second, which I will call "principle-theories." These employ the analytic, not the synthetic, method. The elements which form their basis and starting point are not hypothetically constructed but empirically discovered ones, general characteristics of natural processes, principles that give rise to mathematically formulated criteria which the separate processes or the theoretical representations of them have to satisfy. [Thermodynamics is an example.] ...
>
> The advantages of the constructive theory are completeness, adaptability, and clearness, those of the principle theory are logical perfection and security of the foundations. The theory of relativity belongs to the latter class.

It is worth noting the irony that in the past two decades, just as some have sought a principle explanation of QM, others have sought a constructive explanation of SR [Brown, 2005, Brown and Pooley, 2006]. While we cannot go into detail on such matters, we note that reasons for seeking a principle explanation of QM include not just the ad hoc nature of the postulates, but the fact that there is no agreement on "constructive interpretations," in part because they do nothing but recover what is already in textbook QM, and therefore lead to no new physics or unification. Indeed, non-local interpretations of QM only make unification with SR more problematic.

For those who believe the fundamental explanation for QM phenomena must be constructive, at least in the sense envisioned by Einstein above, none of the mainstream interpretations neatly fit the bill. Not only do most interpretations entail some form of QM holism, contextuality, and/or non-locality, the remainder invoke priority monism and/or multiple branches or outcomes. The problem with attempting a constructive account of QM is, as articulated by [Van Camp, 2011], "Constructive interpretations are attempted, but they are not unequivocally constructive in any traditional sense." Thus, [Van Camp, 2011] states:

> The interpretive work that must be done is less in coming up with a constructive theory and thereby explaining puzzling quantum phenomena, but more in explaining why the interpretation counts as explanatory at all given that it must sacrifice some key aspect of the traditional understanding of causal-mechanical explanation.

It seems clear all of this would be anathema to Einstein and odious with respect to constructive explanation, especially if say, the kinetic theory of gases is the paradigm example of constructive explanation. Thus, for many it seems wise to at least attempt a principle explanation of QM, as sought by QIT. The problem with QIT's attempts is noted by [Van Camp, 2011]:

> However, nothing additional has been shown to be incorporated into an information-theoretic reformulation of QM beyond what is contained in QM itself. It is hard to see how it could offer more unification of the phenomena than QM already does since they are equivalent, and so it is not offering any explanatory value on this front.

Nonetheless, QIT continues to seek "the *reconstruction* of quantum theory" via a constraint-based/principle approach [Chiribella and Spekkens, 2016]. Indeed, QIT has produced several different sets of axioms, postulates, and "physical requirements" in terms of quantum information, which all reproduce quantum theory. Along those lines, [Bub, 2004, Bub, 2012, Bub, 2016] has asked, "why is the world quantum and not classical, and why is it quantum rather than superquantum, i.e., why the Tsirelson bound for quantum correlations?"

Despite all the success of QIT, the community does not find any of the reconstructions compelling. [Cuffaro, 2017], for example, argues that information causality needs to be justified in some physical sense. And, as Hardy states, "When I started on this, what I wanted to see was two or so obvious, compelling axioms that would give you quantum theory and which no one would argue with" [Ball, 2017]. Fuchs quotes Wheeler, "If one really understood the central point and its necessity in the construction of the world, one ought to state it in one clear, simple sentence" [Fuchs and Stacey, 2016, p. 302]. Asked if he had such a sentence, Fuchs responded, "No, that's my big failure at this point" [Fuchs and Stacey, 2016, p. 302]. As we will show, the same principle responsible for the kinematic structure of SR is also responsible for the qubit Hilbert space structure at the foundation of Hardy's and Dakic & Brukner's reconstructions of quantum theory [Hardy, 2016, Dakic and Brukner, 2009], which uniquely produces the Tsirelson bound [Cirel'son, 1980, Landau, 1987, Khalfin and Tsirelson, 1992], viz., "no preferred reference frame" (NPRF, aka the relativity principle) (Figure 1). That is, NPRF applied to the measurement of the speed of light c gives the light postulate and leads to the geometry of Minkowski spacetime whence the Lorentz transformations of SR, while NPRF applied to the measurement of Planck's constant h gives "average-only" projection and leads to the qubit Hilbert space structure whence the QIT reconstructions of QM.

The term "reference frame" has many meanings in physics related to microscopic and macroscopic phenomena, Galilean versus Lorentz transformations, relatively moving observers, etc. The difference between Galilean and Lorentz transformations resides in the fact that the speed of light is finite, so NPRF entails the light postulate of SR [Serway and Jewett, 2019, Knight, 2008], i.e., that everyone measure the same speed of light c, regardless of their motion relative to the source. If there was only one reference frame for a source in which the speed of light equaled the prediction from Maxwell's equations ($c = \frac{1}{\sqrt{\mu_o \epsilon_o}}$), then that would certainly constitute a preferred reference frame. Essentially, Einstein merely extended the relativity principle from mechanics to electromagnetism. Herein, we further extend NPRF to include the measurement of another

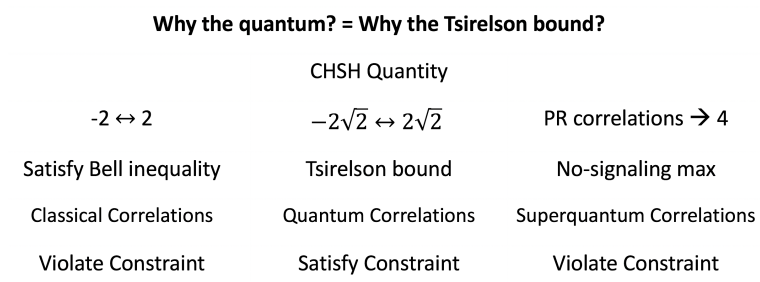


Figure 1: **Answer to Bub's question, "Why the Tsirelson bound?"** The "constraint" is conservation per no preferred reference frame.

fundamental constant of nature, Planck's constant h $(= 2\pi\hbar)$.

As Steven Weinberg points out, measuring an electron's spin via Stern-Gerlach (SG) magnets constitutes the measurement of "a universal constant of nature, Planck's constant" [Weinberg, 2017] (Figure 2). So if NPRF applies equally here, everyone must measure the same value for Planck's constant h regardless of their SG magnet orientations relative to the source, which like the light postulate is an "empirical discovered" fact. By "relative to the source" of a pair of spin-entangled particles, we might mean relative "to the vertical in the plane perpendicular to the line of flight of the particles" [Mermin, 1981, p. 943] ($\hat{z}$ in Figure 3, for example). Here the possible spin outcomes $\pm\frac{\hbar}{2}$ represent a fundamental (indivisible) unit of information per Dakic and Brukner's first axiom in their reconstruction of quantum theory, "An elementary system has the information carrying capacity of at most one bit" [Dakic and Brukner, 2009]. Thus, different SG magnet orientations relative to the source constitute different "reference frames" in QM just as different velocities relative to the source constitute different "reference frames" in SR. Borrowing from [Einstein, 1936], NPRF might be stated:

> No one's "sense experiences," to include measurement outcomes, can provide a privileged perspective on the "real external world."

This is consistent with the notion of symmetries per [Hicks, 2019]:

> There are not two worlds in one of which I am here and in the other I am three feet to the left, with everything else similarly shifted. Instead, there is just this world and two mathematical descriptions of it. The fact that those descriptions put the origin at different places does not indicate any difference between the worlds, as the origin in our mathematical description did not correspond to anything in the world anyway. The symmetries tell us what structure the world does not have.

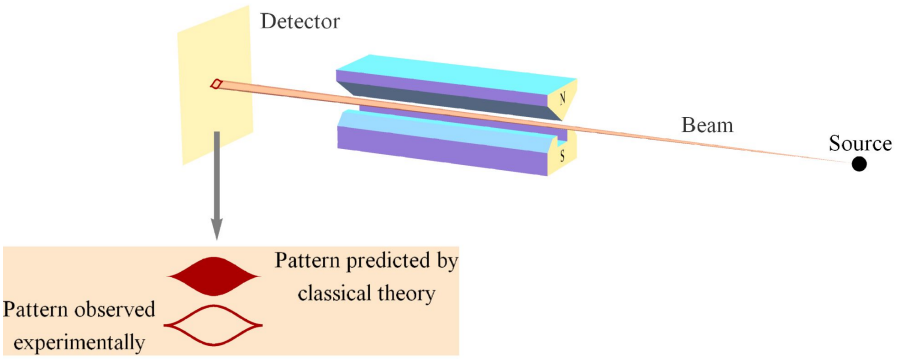


Figure 2: A Stern-Gerlach (SG) spin measurement showing the two possible outcomes, up $(+\frac{\hbar}{2})$ and down $(-\frac{\hbar}{2})$ or $+1$ and -1, for short. The important point to note here is that the classical analysis predicts all possible deflections, not just the two that are observed. The difference between the classical prediction and the quantum reality uniquely distinguishes the quantum joint distribution from the classical joint distribution for the Bell spin states [Garg and Mermin, 1982].

That is, there is just one "real external world" harboring many, but always equal perspectives as far as the physics is concerned [Silberstein and Stuckey, 2020].

We have shown elsewhere that the quantum correlations and quantum states corresponding to the Bell states, which uniquely produce the Tsirelson bound for the Clauser–Horne–Shimony–Holt (CHSH) quantity, can be derived from conservation per NPRF [Stuckey et al., 2019]. Thus, Bell state entanglement is ultimately grounded in NPRF just as SR [Stuckey et al., 2020]. As summarized in Figure 1, the quantum correlations responsible for the Tsirelson bound satisfy conservation per NPRF while both classical and superquantum correlations can violate this constraint. Therefore a principle explanation of Bell state entanglement and the Tsirelson bound that be stated in "one clear, simple sentence" is "conservation per no preferred reference frame" (Figure 1).

What qualifies as a principle explanation versus constructive turns out to be a fraught and nuanced question [Felline, 2011] and we do not want to be sidetracked on that issue as such. Let us therefore state explicitly that what makes our explanation a principle one is that it is grounded directly in phenomenology, it is an adynamical and acausal explanation that involves adynamical global constraints as opposed to dynamical laws or causal mechanisms, and it is unifying with respect to QM and SR.

Let us also note that while contrary to certain others [Brown, 2005, Brown and Pooley, 2006, Norton, 2008, Menon, 2019], we are arguing that conservation per NPRF need not ever be discharged by a constructive explanation or interpretation. This is at least partially distinct from the question in SR

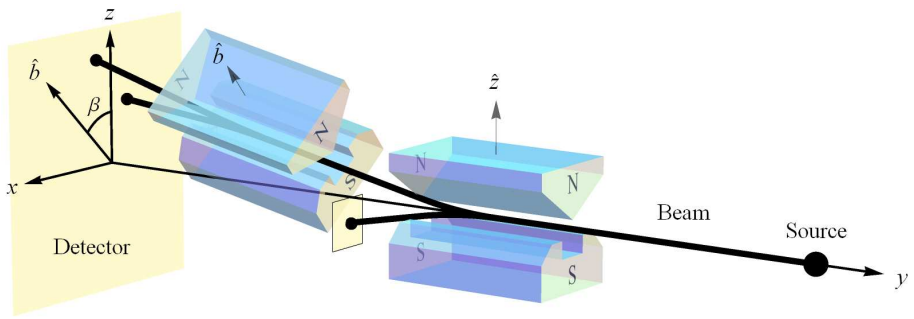


Figure 3: In this set up, the first SG magnets (oriented at $\hat{z}$) are being used to produce an initial state $|\psi\rangle = |u\rangle$ for measurement by the second SG magnets (oriented at $\hat{b}$).

for example, of whether facts about physical geometry are grounded in facts about dynamical fields or vice-versa. Furthermore, this principle explanation is consistent with any number of "constructive interpretations" of QM. For example, this principle explanation avoids the complaints about Bub's proposed principle explanation of QM leveled by [Felline, 2018]. That is, the principle being posited herein does not require a solution to the measurement problem nor again does it necessarily beg for a constructive counterpart.

In Section 2 we provide a quick review of length contraction, time dilation, the relativity of simultaneity, and Lorentz transformations per SR. In Section 3 we review how the qubit Hilbert space structure follows from NPRF [Silberstein et al., 2021] and how that leads to conservation per NPRF responsible for Bell state entanglement [Stuckey et al., 2020] whence the Tsirelson bound [Stuckey et al., 2019]. In Section 4 we argue that principle explanation for these mysteries suffices despite the fact that there is no constructive counterpart.

2 NPRF and Special Relativity

Suppose there are three women moving together at $0.6c$ with respect to two men. The men and women agree on the details of the following four Events (men's coordinates are lower case and women's coordinates are upper case):

- Event 1: Joe meets Sara at $X_1 = x_1 = 0$, $T_1 = t_1 = 0$.
- Event 2: Bob meets Kim at $X_2 = 1250$km, $T_2 = -0.0025$s, $x_2 = 1000$km, $t_2 = 0$.
- Event 3: Bob meets Alice at $X_3 = 800$km, $T_3 = 0$, $x_3 = 1000$km, $t_3 = 0.002$s.

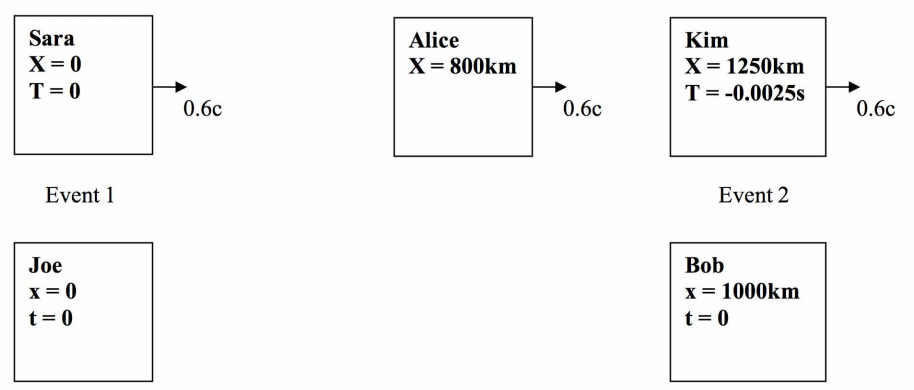


Figure 4: Events 1 and 2 are simultaneous for the men and are spaced at a distance of 1000km. The women say the distance between Sara and Kim is 1250km. Thus, the men say the women's meter sticks are short.

- Event 4: Bob meets Sara at $X_4 = 0$, $T_4 = 0.00\overline{44}$s, $x_4 = 1000$km, $t_4 = 0.00\overline{55}$s.

The lower-case and upper-case coordinates for each Event are related by Lorentz transformations with $\gamma = 1.25$. Here is the story according to the men.

The women are moving in the positive x direction at $0.6c$. Events 1 and 2 are simultaneous ($t_1 = t_2 = 0$), so the distance between Sara and Kim is $x_2 = 1000$km. The women say the distance between Sara and Kim is $X_2 = 1250$km, so their proper distance has been length contracted by γ (Figure 4). Event 4 happens $t_4 = 0.00\overline{55}$s after Events 1 and 2, but Sara's clock has only ticked off $T_4 = 0.00\overline{44}$s, so her proper time has been dilated by a factor of γ (Figure 5). Therefore, the men say the women's meter sticks are short (length contraction) and the women's clocks are running slow (time dilation). Here is the story according to the women.

The men are moving in the negative X direction at $0.6c$. Events 1 and 3 are simultaneous ($T_1 = T_3 = 0$), not Events 1 and 2 as the men claim (relativity of simultaneity). Thus, the distance between Joe and Bob is $X_3 = 800$km, not $x_3 = 1000$km as the men claim (Figure 6). Again, the proper distance has been length contracted by γ. Event 3 happens 0.0025s after Event 2, but Bob's clock has only ticked off $t_3 = 0.002$s, so his proper time has been dilated by a factor of γ (Figure 6). Therefore, the women say the men's meter sticks are short and the men's clocks are running slow.

In summary, NPRF gives the postulates of SR whence the Lorentz transformations, time dilation, length contraction, and the relativity of simultaneity. Since Alice and Bob always measure the same speed of light c regardless of their relative motion per NPRF, Alice says Bob's temporal and spatial measurements need to be corrected per time dilation and length contraction while Bob says

Figure 5: According to the men, Event 4 happens $t_4 = 0.00\overline{55}$s after Events 1 and 2, but Sara's clock has only ticked off $T_4 = 0.00\overline{44}$s. Thus, the men say the women's clocks are running slow.

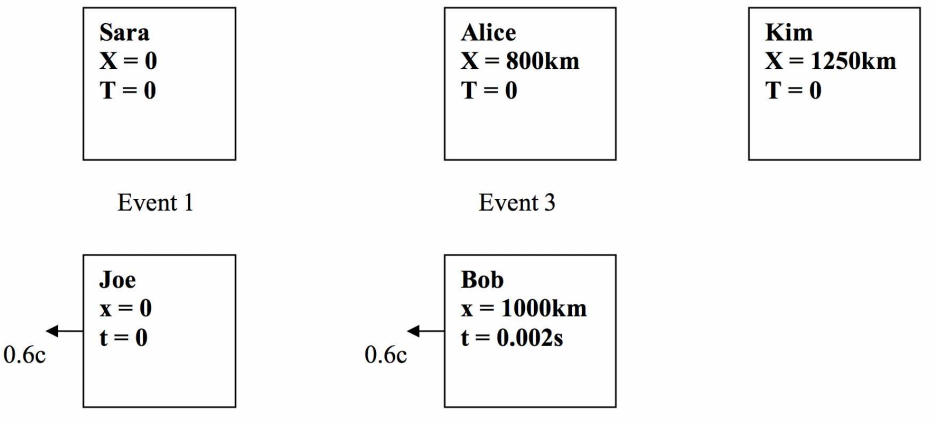


Figure 6: According to the women, Events 1 and 3 are simultaneous not Events 1 and 2 as the men claim (relativity of simultaneity). Thus, the distance between Joe and Bob is $X_3 = 800$km, not $x_3 = 1000$km as the men claim, i.e., the women say the men's meter sticks are short. Also, Event 3 happens 0.0025s after Event 2, but Bob's clock has only ticked off $t_3 = 0.002$s, so the women say the men's clocks are running slow.

the same thing about Alice's measurements. But, if NPRF is true and fundamental, then neither need to be corrected (relativity of simultaneity). Thus, the mysteries of length contraction and time dilation in SR ultimately reside in NPRF starting with the fact that everyone measures the same value for the fundamental constant c. Now let us relate this mystery to the mystery of Bell state entanglement in QM.

3 NPRF and Quantum Mechanics

3.1 "Average-Only" Projection per NPRF

While we will refer explicitly to SG spin measurements, this can be understood to represent any measurement with a binary outcome in the symmetry plane. The other possible outcome is normal to the symmetry plane, as in "V" $(+1)$ or "H" (-1) outcomes with photons and polarizers, where one has "intensity of the transmitted beam" rather than "projection of the transmitted vector" [Stuckey et al., 2019]. In either case, the outcome represents the invariant measure of the fundamental unit of action h with respect to the SO(3) transformations between QM reference frames, as in all quantum exchanges [Silberstein and Stuckey, 2020]. SO(3) with Lorentz boosts then complete the restricted Lorentz transformation group between reference frames. As shown explicitly by Dakic & Brukner [Dakic and Brukner, 2009], the SO(3) transformation group uniquely identifies the fundamental probability structure of QM amid those of classical probability theory and higher-dimensional generalized probability theories [Silberstein et al., 2021].

If we create a preparation state oriented along the positive z axis as in Figure 3, i.e., $|\psi\rangle = |u\rangle$, our spin angular momentum is $\vec{S} = +1\hat{z}$ (in units of $\frac{\hbar}{2} = 1$). Now proceed to make a measurement with the SG magnets oriented at $\hat{b}$ making an angle β with respect to $\hat{z}$ (Figure 3). According to classical physics, we expect to measure $\vec{S} \cdot \hat{b} = \cos(\beta)$ (Figure 7), but we cannot measure anything other than ± 1 due to NPRF (contra the prediction by classical physics), so we see that NPRF answers Wheeler's "Really Big Question," "Why the quantum?" [Wheeler, 1986, Barrow et al., 2004] in "one clear, simple sentence" to convey "the central point and its necessity in the construction of the world." As a consequence, we can only recover $\cos(\beta)$ *on average* (Figure 8), i.e., NPRF dictates "average-only" projection

$$(+1)P(+1 \mid \beta) + (-1)P(-1 \mid \beta) = \cos(\beta) \tag{1}$$

Solving simultaneously with $P(+1 \mid \beta) + P(-1 \mid \beta) = 1$, we find that

$$P(+1 \mid \beta) = \cos^2\left(\frac{\beta}{2}\right) \tag{2}$$

and

$$P(-1 \mid \beta) = \sin^2\left(\frac{\beta}{2}\right) \tag{3}$$

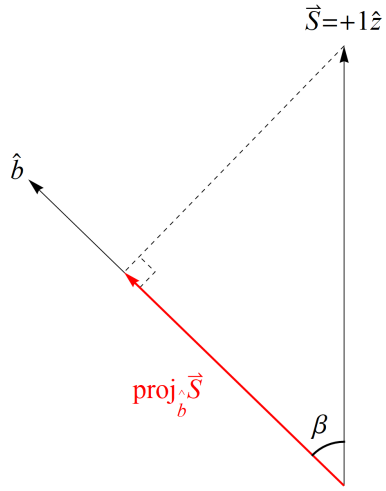


Figure 7: The spin angular momentum of Bob's particle $\vec{S}$ projected along his measurement direction $\hat{b}$. This does *not* happen with spin angular momentum due to NPRF.

When talking about the longitudinal outcomes [Dehlinger and Mitchell, 2002] ("click" or "no click"), we have

$$P(V \mid \beta) = \cos^2(\beta) \tag{4}$$

and

$$P(H \mid \beta) = \sin^2(\beta) \tag{5}$$

so that our average outcome at β (orientation of polarizer with respect to initial polarization state) is given by

$$(+1)\cos^2(\beta) + (-1)\sin^2(\beta) = \cos^2(\beta) - \sin^2(\beta) \tag{6}$$

This is the naively expected Malus law per classical physics for the intensity of electromagnetic radiation transmitted through a polarizer if "pass" is $+1$ and "no pass" is -1 (instead of 0). As with the transverse mode NPRF rules out "fractional outcomes," again contra the prediction by classical physics, so the classical result obtains only on average when $\beta \neq 0$. This explains the ineluctably probabilistic nature of QM, as pointed out by Mermin [Mermin, 2019, p. 10]:

> Quantum mechanics is, after all, the first physical theory in which probability is explicitly not a way of dealing with ignorance of the precise values of existing quantities.

Of course, these "average-only" results due to "no fractional outcomes per NPRF" hold precisely for the qubit Hilbert space structure of QM.

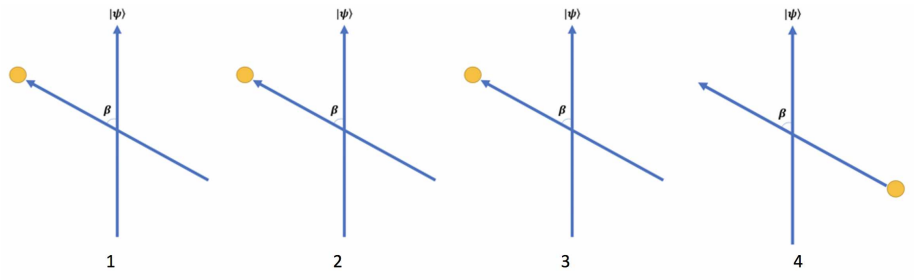


Figure 8: An ensemble of 4 SG measurement trials for $\beta = 60°$ in Figure 3. The tilted blue arrow depicts an SG measurement orientation and the vertical arrow represents our preparation state $|\psi\rangle = |u\rangle$. The yellow dots represent the two possible measurement outcomes for each trial, up (located at arrow tip) or down (located at bottom of arrow). The expected projection result of $\cos(\beta)$ cannot be realized because the measurement outcomes are binary (quantum) with values of $+1$ (up) or -1 (down) per NPRF. Thus, we have "average-only" projection for all 4 trials (three up outcomes and one down outcome average to $\cos(60°) = \frac{1}{2}$). That is, the *average* of the ± 1 outcomes equals the projection of the initial spin angular momentum vector $\vec{S} = +1\hat{z}$ in the measurement direction $\hat{b}$, i.e., $\vec{S} \cdot \hat{b} = \cos(60°) = \frac{1}{2}$.

Let's explicitly review the qubit Hilbert space structure represented by the Pauli spin matrices. In the eigenbasis of σ_z the Pauli spin matrices are

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -\mathbf{i} \\ \mathbf{i} & 0 \end{pmatrix}, \quad \text{and} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

where $\mathbf{i} = \sqrt{-1}$. All spin matrices have the same ± 1 eigenvalues (measurement outcomes), which reflects the fact that there are no fractional outcomes per NPRF. We denote the corresponding eigenvectors (eigenstates) as $|u\rangle$ and $|d\rangle$ for spin up $(+1)$ and spin down (-1), respectively. Using the Pauli spin matrices supra with $|u\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $|d\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, we see that $\sigma_z|u\rangle = |u\rangle$, $\sigma_z|d\rangle = -|d\rangle$, $\sigma_x|u\rangle = |d\rangle$, $\sigma_x|d\rangle = |u\rangle$, $\sigma_y|u\rangle = \mathbf{i}|d\rangle$, and $\sigma_y|d\rangle = -\mathbf{i}|u\rangle$. If we change the orientation of a vector from right pointing (ket) to left pointing (bra) or vice-versa, we transpose and take the complex conjugate. For example, if $|A\rangle = \mathbf{i}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{i}|u\rangle$, then $\langle A| = -\mathbf{i}\begin{pmatrix} 1 & 0 \end{pmatrix} = -\mathbf{i}\langle u|$. Therefore, any spin matrix can be written as $(+1)|u\rangle\langle u| + (-1)|d\rangle\langle d|$ where $|u\rangle$ and $|d\rangle$ are their up and down eigenstates, respectively. A qubit is then constructed from this two-level quantum system, i.e., $|\psi\rangle = c_1|u\rangle + c_2|d\rangle$ where $|c_1|^2 + |c_2|^2 = 1$.

An arbitrary spin measurement σ in the $\hat{b}$ direction is given by the spin matrices

$$\sigma = \hat{b} \cdot \vec{\sigma} = b_x\sigma_x + b_y\sigma_y + b_z\sigma_z \tag{7}$$

Again, preparation states $|\psi\rangle$ are created from linear combinations of the Pauli spin eigenstates. The average outcome (all we can obtain per NPRF) for a measurement σ on state $|\psi\rangle$ is given by

$$\langle\sigma\rangle := \langle\psi|\sigma|\psi\rangle \tag{8}$$

For example, in Figure 3 we have $|\psi\rangle = |u\rangle$ (prepared by the first SG magnets) and $\sigma = \sin(\beta)\sigma_x + \cos(\beta)\sigma_z$ (per the second SG magnets), so $\langle\sigma\rangle = \cos(\beta)$ in accord with Eq. (1).

Finally, the probability of obtaining a $+1$ or -1 result for σ is just

$$P(+1 \mid \beta) = |\langle\psi|\tilde{u}\rangle|^2 = \cos^2\left(\frac{\beta}{2}\right) \tag{9}$$

and

$$P(-1 \mid \beta) = |\langle\psi|\tilde{d}\rangle|^2 = \sin^2\left(\frac{\beta}{2}\right) \tag{10}$$

where $|\tilde{u}\rangle$ and $|\tilde{d}\rangle$ are the eigenvectors of σ and $\frac{\beta}{2}$ is the angle between $|\psi\rangle$ and $|\tilde{u}\rangle$ in Hilbert space. This agrees with the result from NPRF in Eqs. (2) & (3). Thus, we see how the principle of NPRF underwrites the QM operational structure for qubits and, therefore, the QIT reconstructions of QM built upon the qubit. In the following, we briefly review the SU(2)/SO(3) transformation property for qubits via their bipartite entanglement in the Bell spin states and show how this Hilbert space structure also follows from NPRF.

3.2 "Average-Only" Conservation per NPRF

When considering two-particle states, we will use the juxtaposed notation for our spin states and matrices. Thus, $\sigma_x\sigma_z|ud\rangle = -|dd\rangle$ and $\sigma_x\sigma_y|ud\rangle = -\mathbf{i}|du\rangle$, for example. Essentially, we are simply ignoring the tensor product sign $\otimes$, so that $(\sigma_x \otimes \sigma_z)\,||u\rangle \otimes |d\rangle\rangle = \sigma_x\sigma_z|ud\rangle$. It is still easy to see which spin matrix is acting on which Hilbert space vector via the juxtaposition. The Bell states are

$$\begin{aligned}
|\psi_-\rangle &= \frac{|ud\rangle - |du\rangle}{\sqrt{2}} \\
|\psi_+\rangle &= \frac{|ud\rangle + |du\rangle}{\sqrt{2}} \\
|\phi_-\rangle &= \frac{|uu\rangle - |dd\rangle}{\sqrt{2}} \\
|\phi_+\rangle &= \frac{|uu\rangle + |dd\rangle}{\sqrt{2}}
\end{aligned} \tag{11}$$

in the eigenbasis of σ_z. The first state $|\psi_-\rangle$ is called the "spin singlet state" and it represents a total conserved spin angular momentum of zero ($S = 0$) for the two particles involved. The other three states are called the "spin triplet

states" and they each represent a total conserved spin angular momentum of one ($S = 1$, in units of $\hbar = 1$ for spin-$\frac{1}{2}$ particles). In all four cases, the entanglement represents the conservation of spin angular momentum for the process creating the state.

If Alice is making her spin measurement σ_1 in the $\hat{a}$ direction and Bob is making his spin measurement σ_2 in the $\hat{b}$ direction (Figure 9), we have

$$\begin{aligned} \sigma_1 &= \hat{a} \cdot \vec{\sigma} = a_x\sigma_x + a_y\sigma_y + a_z\sigma_z \\ \sigma_2 &= \hat{b} \cdot \vec{\sigma} = b_x\sigma_x + b_y\sigma_y + b_z\sigma_z \end{aligned} \tag{12}$$

The correlation functions are given by [Stuckey et al., 2020]

$$\begin{aligned} \langle\psi_-|\sigma_1\sigma_2|\psi_-\rangle &= \quad -a_xb_x - a_yb_y - a_zb_z \\ \langle\psi_+|\sigma_1\sigma_2|\psi_+\rangle &= \quad a_xb_x + a_yb_y - a_zb_z \\ \langle\phi_-|\sigma_1\sigma_2|\phi_-\rangle &= \quad -a_xb_x + a_yb_y + a_zb_z \\ \langle\phi_+|\sigma_1\sigma_2|\phi_+\rangle &= \quad a_xb_x - a_yb_y + a_zb_z \end{aligned} \tag{13}$$

The spin singlet state is invariant under all three SU(2) transformations meaning we obtain opposite outcomes ($\frac{1}{2}$ ud and $\frac{1}{2}$ du) for SG magnets at any $\hat{a} = \hat{b}$ (Figure 9) and a correlation function of $-\cos(\theta)$ in any plane of physical space, where θ is the angle between $\hat{a}$ and $\hat{b}$ (Eq. (13)). We see that the conserved spin angular momentum ($S = 0$), being directionless, is conserved in any plane of physical space. Again, $\hat{a} = \hat{b}$ means Alice and Bob are in the same reference frame.

The invariance of each of the spin triplet states under its respective SU(2) transformation in Hilbert space represents the SO(3) invariant conservation of spin angular momentum $S = 1$ for each of the planes xz ($|\phi_+\rangle$), yz ($|\phi_-\rangle$), and xy ($|\psi_+\rangle$) in physical space. Specifically, when the SG magnets are aligned (the measurements are being made in the same reference frame) anywhere in the respective plane of symmetry the outcomes are always the same ($\frac{1}{2}$ uu and $\frac{1}{2}$ dd). It is a planar conservation and our experiment would determine which plane. If you want to model a conserved $S = 1$ for some other plane, you simply create a superposition, i.e., expand in the spin triplet basis. And in that plane, you're right back to the mystery of Bell state entanglement per conserved spin angular momentum via a correlation function of $\cos(\theta)$, as with any of the spin triplet states (Eq. (13)). We will now explain how the spin singlet state correlation function follows from NPRF (the spin triplet state correlation function is analogous).

That we have opposite outcomes when Alice and Bob are in the same reference frame is not difficult to understand via conservation of spin angular momentum, because Alice and Bob's measured values of spin angular momentum cancel directly when $\hat{a} = \hat{b}$ (Figure 9). But, when Bob's SG magnets are rotated by θ relative to Alice's SG magnets, we need to clarify the situation.

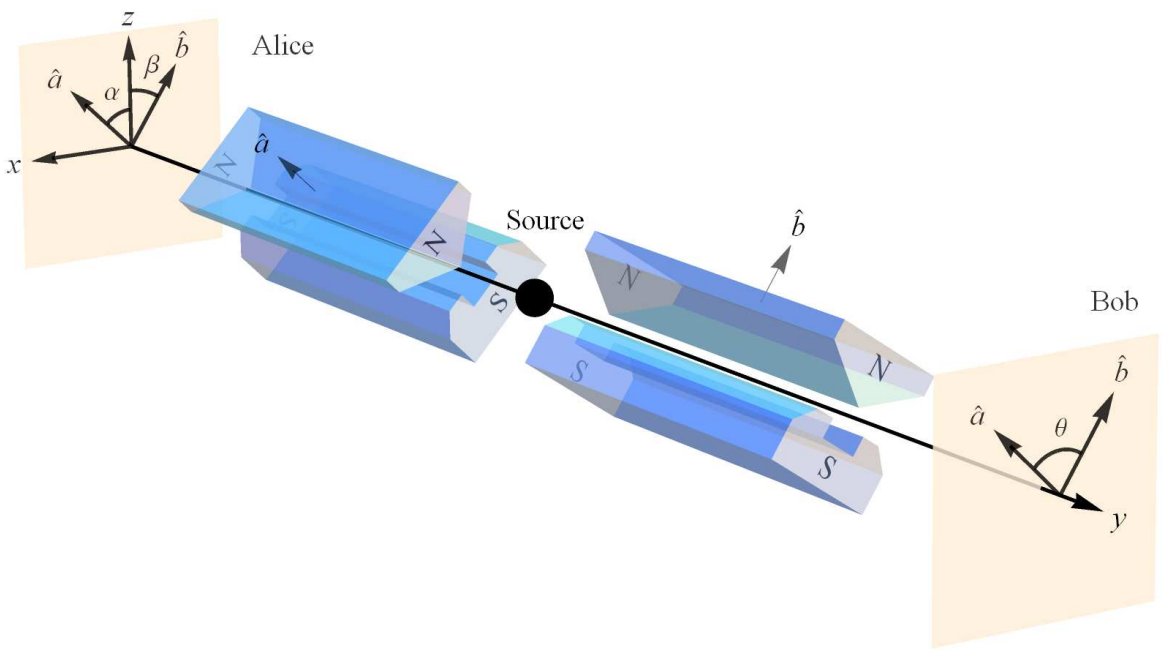


Figure 9: Alice and Bob making spin measurements on a pair of spin-entangled particles with their Stern-Gerlach (SG) magnets and detectors in the xz-plane. Here Alice and Bob's SG magnets are not aligned so these measurements represent different reference frames. Since their outcomes satisfy NPRF in all reference frames and satisfy explicit conservation of spin angular momentum in the same reference frame, they can only satisfy conservation of spin angular momentum on *average* in different reference frames.

We have two subsets of data, Alice's set (with SG magnets at angle α) and Bob's set (with SG magnets at angle β). They were collected in N pairs (data events) with Bob's(Alice's) SG magnets at $\alpha - \beta = \theta$ relative to Alice's(Bob's). We want to compute the correlation function for these N data events which is

$$\langle \alpha, \beta \rangle = \frac{(+1)_A(-1)_B + (+1)_A(+1)_B + (-1)_A(-1)_B + ...}{N} \tag{14}$$

Now partition the numerator into two equal subsets per Alice's equivalence relation, i.e., Alice's $+1$ results and Alice's -1 results

$$\langle \alpha, \beta \rangle = \frac{(+1)_A(\sum \text{BA+}) + (-1)_A(\sum \text{BA-})}{N} \tag{15}$$

where $\sum$BA+ is the sum of all of Bob's results (event labels) corresponding to Alice's $+1$ result (event label) and $\sum$BA- is the sum of all of Bob's results (event labels) corresponding to Alice's -1 result (event label). Next, rewrite Eq. (15) as

$$\langle \alpha, \beta \rangle = \frac{1}{2}(+1)_A\overline{BA+} + \frac{1}{2}(-1)_A\overline{BA-} \tag{16}$$

with the overline denoting average. Eq. (16) is independent of the formalism of QM, all we have assumed is that Alice and Bob each measure $+1$ and -1 with equal frequency for all measurement settings per NPRF. Notice that to understand the quantum correlation responsible for Bell state entanglement, we need to understand the origins of $\overline{BA+}$ and $\overline{BA-}$ for the Bell states. We now show what that is for the spin singlet state (the spin triplet states are analogous in their respective symmetry planes [Stuckey et al., 2020, Silberstein et al., 2021]).

In classical physics, one would say the projection of the spin angular momentum vector of Alice's particle $\vec{S}_A = +1\hat{a}$ along $\hat{b}$ is $\vec{S}_A \cdot \hat{b} = +\cos(\theta)$ where again θ is the angle between the unit vectors $\hat{a}$ and $\hat{b}$. That's because the prediction from classical physics is that all values between $+1\left(\frac{\hbar}{2}\right)$ and $-1\left(\frac{\hbar}{2}\right)$ are possible outcomes for a spin angular momentum measurement (Figure 2). From Alice's perspective, had Bob measured at the same angle, i.e., $\beta = \alpha$, he would have found the spin angular momentum vector of his particle was $\vec{S}_B = -\vec{S}_A = -1\hat{a}$, so that $\vec{S}_A + \vec{S}_B = \vec{S}_{Total} = 0$. Since he did not measure the spin angular momentum of his particle at the same angle, he should have obtained a fraction of the length of $\vec{S}_B$, i.e., $\vec{S}_B \cdot \hat{b} = -1\hat{a} \cdot \hat{b} = -\cos(\theta)$ (Figure 10). But according to NPRF, Bob only ever obtains $+1$ or -1 just like Alice, so he cannot measure the required fractional outcome to explicitly conserve spin angular momentum per Alice. Therefore, as with the single-particle case, NPRF means that Bob's outcomes must satisfy "average-only" projection (Figure 11), which means

$$\overline{BA+} = -\cos(\theta) \tag{17}$$

Given this constraint per NPRF, as with the single-particle case, we can now use NPRF to find the joint probabilities for Alice and Bob's outcome pairs. Looking at Table 1, the rows and columns all sum to $\frac{1}{2}$ because both Alice and Bob must observe $+1$ half of the time and -1 half of the time per NPRF, which

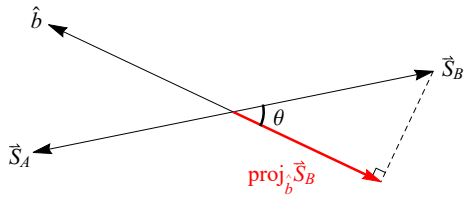


Figure 10: The spin angular momentum of Bob's particle $\vec{S}_B = -\vec{S}_A$ projected along his measurement direction $\hat{b}$. This does *not* happen with spin angular momentum.

		Bob		
		$+1$	-1	Total
Alice	$+1$	$P(+1,+1 \mid \theta)$	$P(+1,-1 \mid \theta)$	$1/2$
	-1	$P(+1,-1 \mid \theta)$	$P(-1,-1 \mid \theta)$	$1/2$
	Total	$1/2$	$1/2$	1

Table 1: **Joint probabilities for Alice and Bob's outcome pairs for the entangled particle experiment in Figure 9.** The table is symmetric due to NPRF.

also asserts that the table is symmetric so that $P(-1,+1 \mid \theta) = P(+1,-1 \mid \theta)$. The average of Bob's outcomes given that Alice observes a $+1$ is

$$\overline{BA+} = 2P(+1,+1 \mid \theta)(+1) + 2P(+1,-1 \mid \theta)(-1) = -\cos(\theta) \tag{18}$$

using conservation per NPRF. Together with the constraints on the rows/columns

$$P(+1,+1 \mid \theta) + P(+1,-1 \mid \theta) = \frac{1}{2}$$
$$P(+1,-1 \mid \theta) + P(-1,-1 \mid \theta) = \frac{1}{2},$$

we can uniquely solve for the joint probabilities

$$P(+1,+1 \mid \theta) = P(-1,-1 \mid \theta) = \frac{1}{2}\sin^2\left(\frac{\theta}{2}\right) \tag{19}$$

and

$$P(+1,-1 \mid \theta) = P(-1,+1 \mid \theta) = \frac{1}{2}\cos^2\left(\frac{\theta}{2}\right). \tag{20}$$

Now we can use these to compute $\overline{BA-}$

$$\overline{BA-} = 2P(-1,+1 \mid \theta)(+1) + 2P(-1,-1 \mid \theta)(-1) = \cos(\theta) \tag{21}$$

Figure 11: **Average View for the Spin Singlet State**. Reading from left to right, as Bob rotates his SG magnets relative to Alice's SG magnets for her +1 outcome, the average value of his outcome varies from −1 (totally down, arrow bottom) to 0 to +1 (totally up, arrow tip). This obtains per conservation of spin angular momentum on average in accord with no preferred reference frame. Bob can say exactly the same about Alice's outcomes as she rotates her SG magnets relative to his SG magnets for his +1 outcome. That is, their outcomes can only satisfy conservation of spin angular momentum *on average* in different reference frames, because they only measure ±1, never a fractional result. Thus, just as with the light postulate of SR, we see that no preferred reference frame leads to a counterintuitive result. Here it requires quantum outcomes $\pm 1\left(\frac{\hbar}{2}\right)$ for all measurements and that leads to the mystery of "average-only" conservation.

Using Eqs. (18) & (21) in Eq. (16) we obtain

$$\langle \alpha, \beta \rangle = \frac{1}{2}(+1)_A(-\cos(\theta)) + \frac{1}{2}(-1)_A(\cos(\theta)) = -\cos(\theta) \tag{22}$$

which is precisely the correlation function for the spin singlet state.

There are two important points to be made here. First, NPRF is just the statement of an "empirically discovered" fact, i.e., Alice and Bob both always measure ±1. Second, it is simply a mathematical fact that the "average-only" conservation of Eqs. (18) & (21) yields the quantum correlation functions of Eq. (13). In other words, to paraphrase Einstein, "we have an empirically discovered principle that gives rise to mathematically formulated criteria which the separate processes or the theoretical representations of them have to satisfy." That is why this principle account of quantum entanglement provides "logical perfection and security of the foundations." Thus, we see how quantum entanglement follows from NPRF applied to the measurement of h in precisely the same manner that time dilation and length contraction follow from NPRF applied to the measurement of c. And, just like in SR, Bob could partition the data according to his equivalence relation (per his reference frame) and claim that it is Alice who must average her results (obtained in her reference frame) to conserve spin angular momentum (Table 2).

Special Relativity	**Quantum Mechanics**
Empirical Fact: Alice and Bob both measure c, regardless of their motion relative to the source	Empirical Fact: Alice and Bob both measure $\pm 1\left(\frac{\hbar}{2}\right)$, regardless of their SG orientation relative to the source
Alice(Bob) says of Bob(Alice): Must correct time and length measurements	Alice(Bob) says of Bob(Alice): Must average results
NPRF: Relativity of simultaneity	NPRF: Relativity of data partition

Table 2: **Comparing SR with QM according to no preferred reference frame (NPRF)**.

4 Principle versus Constructive Explanation for Bell State Entanglement

As we saw in Section 2 for SR, if Alice is moving at velocity $\vec{V}_a$ relative to a light source, then she measures the speed of light from that source to be c ($= \frac{1}{\sqrt{\mu_o \epsilon_o}}$, as predicted by Maxwell's equations). If Bob is moving at velocity $\vec{V}_b$ relative to that same light source, then he measures the speed of light from that source to be c. Here "reference frame" refers to the relative motion of the observer and source, so all observers who share the same relative velocity with respect to the source occupy the same reference frame. NPRF in this context means all measurements produce the same outcome c.

As a consequence of this constraint we have time dilation and length contraction, which are then reconciled per NPRF via the relativity of simultaneity. That is, Alice and Bob each partition spacetime per their own equivalence relations (per their own reference frames), so that equivalence classes are their own surfaces of simultaneity. If Alice's equivalence relation over the spacetime events yields the "true" partition of spacetime, then Bob must correct his lengths and times per length contraction and time dilation. Of course, the relativity of simultaneity says that Bob's equivalence relation is as valid as Alice's per NPRF.

This is completely analogous to QM, where Alice and Bob each partition the data per their own equivalence relations (per their own reference frames), so that equivalence classes are their own $+1$ and -1 data events. If Alice's equivalence relation over the data events yields the "true" partition of the data, then Bob must correct (average) his results per average-only conservation. Of course, NPRF says that Bob's equivalence relation is as valid as Alice's, which we might call the "relativity of data partition" (Table 2).

Thus, the mysteries of SR (time dilation and length contraction) ultimately follow from the same principle as Bell state entanglement, i.e., no preferred reference frame. So, if one accepts SR's principle explanation of time dilation and length contraction, then they should have no problem accepting conservation per NPRF as a principle explanation of Bell state entanglement. Thus, the relativity principle (NPRF) is a unifying principle for (non-relativistic) QM and

SR, thereby addressing the desideratum of QIT in general and answering Bub's question specifically (Figure 1).

Despite the fact that this principle explanation supplies a unifying framework for both QM and SR, some might demand a constructive explanation with its corresponding "knowledge of how things in the world work, that is, of the mechanisms (often hidden) that produce the phenomena we want to understand" [Salmon, 1993, p. 15]. This is "the causal/mechanical view of scientific explanation" per [Salmon, 1993, p. 15]. Thus, as with SR, not everyone will consider our principle account to be explanatory since, "By its very nature such a theory-of-principle explanation will have nothing to say about the reality behind the phenomenon" [Balashov and Janssen, 2003, p. 331]. As stated by [Brown and Pooley, 2006, p. 76]:

> What has been shown is that rods and clocks must behave in quite particular ways in order for the two postulates to be true together. But this hardly amounts to an explanation of such behaviour. Rather things go the other way around. It is because rods and clocks behave as they do, in a way that is consistent with the relativity principle, that light is measured to have the same speed in each inertial frame.

In other words, the assumption is that the true or fundamental "explanation" of Bell state entanglement must be a constructive one in the sense of adverting to causal mechanisms like fundamental physical entities such as particles or fields and their dynamical equations of motion. Notice that while our account of SR is in terms of fundamental principle explanation, that does not necessarily make it a "geometric" interpretation of SR. For example, nothing we've said commits us to the claim that if one were to remove all the matter-energy out of the universe there would be some geometric structure remaining such as Minkowski spacetime. Furthermore, there is nothing inherently geometric about our principle explanation of Bell state entanglement in particular or of NPRF in general.

Of course we do not have a no-go argument that our principle explanation will never be subsumed by a constructive one. However, especially in light of the unifying nature of our principle explanation, we think it is worth considering the possibility that principle explanation is fundamental in these cases and perhaps others [Silberstein et al., 2018, Stuckey et al., 2019, Stuckey et al., 2020, Silberstein et al., 2021]. We think this is especially reasonable in light of the current impasse in both QIT-based explanations of QM phenomena and in attempts at constructive interpretations. Essentially, we are in a situation with QM that Einstein found himself in with SR [Einstein, 1949, pp. 51-52]:

> By and by I despaired of the possibility of discovering the true laws by means of constructive efforts based on known facts. The longer and the more despairingly I tried, the more I came to the conviction that only the discovery of a universal formal principle could lead us to assured results. The example I saw before me was thermodynamics.

Thus we are offering a competing account of quantum entanglement for any interpretation that fundamentally explains entanglement in the constructive sense. As Einstein said, this gives us the advantage of "logical perfection and security of the foundations" as our principle account could be true across a number of different constructive interpretations. And, the principle we offer, NPRF, is a unifying principle for QM and SR that holds throughout physics [Silberstein and Stuckey, 2020]. As Pauli once stated [Heisenberg, 1971, p. 33]:

> 'Understanding' probably means nothing more than having whatever ideas and concepts are needed to recognize that a great many different phenomena are part of a coherent whole.

Per [Hicks, 2019], NPRF is a principle that is accessible ("because it is simple") and whence we can "infer lots of truths." Inferring "lots of truths" implies a unifying principle is superior to its subsumed constituents, since it implies (at minimum) more truths than any proper subset of its subsumed constituents. The point is, we are hypothesizing that the SO(3) symmetry with average-only conservation as an explanation of Bell state entanglement, and Lorentz symmetry with relativity of simultaneity as an explanation of length contraction and time dilation, are expressions of a deeper truth, NPRF, with seemingly disparate multiple physical consequences. It has been suggested that perhaps other unresolved phenomena in physics might be explained in a similar fashion [Silberstein et al., 2018].

The bottom line is that a compelling constraint (who would argue with conservation per NPRF?) explains Bell state entanglement without any obvious corresponding 'dynamical/causal influence' or hidden variables to account for the results on a trial-by-trial basis. By accepting this principle explanation as fundamental, the lack of a compelling, consensus constructive explanation is not a problem. This is just one of many mysteries in physics created by dynamical and causal biases that can be resolved by constraint-based thinking [Silberstein et al., 2018].

References

[Balashov and Janssen, 2003] Balashov, Y. and Janssen, M. (2003). Presentism and relativity. *British Journal for the Philosophy of Science*, 54:327–346.

[Ball, 2017] Ball, P. (2017). Physicists want to rebuild quantum theory from scratch. *Wired*. https://www.wired.com/story/physicists-want-to-rebuild-quantum-theory-from-scratch/amp.

[Barrow et al., 2004] Barrow, J. D., Davies, P. C. W., and Charles L. Harper, J., editors (2004). *Science and Ultimate Reality: Quantum Theory, Cosmology, and Complexity*. Cambridge university Press, New York.

[Brown, 2005] Brown, H. (2005). *Physical Relativity: Spacetime Structure from a Dynamical Perspective*. Oxford University Press, Oxford, UK.

[Brown and Pooley, 2006] Brown, H. and Pooley, O. (2006). Minkowski spacetime: A glorious non-entity. In Dieks, D., editor, *The Ontology of Spacetime*, page 67. Elsevier, Amsterdam.

[Bub, 2004] Bub, J. (2004). Why the quantum? *Studies in History and Philosophy of Modern Physics*, 35B:241–266. https://arxiv.org/abs/quant-ph/0402149.

[Bub, 2012] Bub, J. (2012). Why the Tsirelson bound? In Hemmo, M. and Ben-Menahem, Y., editors, *The Probable and the Improbable: The Meaning and Role of Probability in Physics*, pages 167–185. Springer, Dordrecht. https://arxiv.org/abs/1208.3744.

[Bub, 2016] Bub, J. (2016). *Bananaworld: Quantum Mechanics for Primates*. Oxford University Press, Oxford, UK.

[Chiribella and Spekkens, 2016] Chiribella, G. and Spekkens, R. (2016). Introduction. In Chiribella, G. and Spekkens, R., editors, *Quantum Theory: Informational Foundations and Foils*, pages 1–18. Springer, Dordrecht.

[Cirel'son, 1980] Cirel'son, B. (1980). Quantum generalizations of Bell's inequality. *Letters in Mathematical Physics*, 4:93–100. https://www.tau.ac.il/ tsirel/download/qbell80.pdf.

[Cuffaro, 2017] Cuffaro, M. E. (2017). Information causality, the Tsirelson bound, and the 'being-thus' of things. http://philsci-archive.pitt.edu/14027/1/tbound.pdf.

[Dakic and Brukner, 2009] Dakic, B. and Brukner, C. (2009). Quantum theory and beyond: Is entanglement special? In Halvorson, H., editor, *Deep Beauty: Understanding the Quantum World through Mathematical Innovation*, pages 365–392. Cambridge University Press. https://arxiv.org/abs/0911.0695.

[Dehlinger and Mitchell, 2002] Dehlinger, D. and Mitchell, M. (2002). Entangled photons, nonlocality, and bell inequalities in the undergraduate laboratory. *American Journal of Physics*, 70(9):903–910.

[Einstein, 1919] Einstein, A. (1919). What is the theory of relativity? *London Times*, pages 53–54.

[Einstein, 1936] Einstein, A. (1936). Physics and reality. *Journal of the Franklin Institute*, 221(3):349–382.

[Einstein, 1949] Einstein, A. (1949). Autobiographical notes. In Schilpp, P. A., editor, *Albert Einstein: Philosopher-Scientist*, pages 3–94. Open Court, La Salle, IL, USA.

[Felline, 2011] Felline, L. (2011). Scientific explanation between principle and constructive theories. *Philosophy of Science*, 78:989–1000.

[Felline, 2018] Felline, L. (2018). Quantum theory is not only about information. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*. https://arxiv.org/abs/1806.05323.

[Fuchs and Stacey, 2016] Fuchs, C. and Stacey, B. (2016). Some negative remarks on operational approaches to quantum theory. In Chiribella, G. and Spekkens, R., editors, *Quantum Theory: Informational Foundations and Foils*, pages 283–305. Springer, Dordrecht.

[Garg and Mermin, 1982] Garg, A. and Mermin, N. (1982). Bell inequalities with a range of violation that does not diminish as the spin becomes arbitrarily large. *Phys. Rev. Lett.*, 49:901–904.

[Hardy, 2016] Hardy, L. (2016). Reconstructing quantum theory. In Chiribella, G. and Spekkens, R., editors, *Quantum Theory: Informational Foundations and Foils*, pages 223–248. Springer, Dordrecht. https://arxiv.org/abs/1303.1538.

[Heisenberg, 1971] Heisenberg, W. (1971). *Physics and Beyond: Encounters and Conversations*. Harper & Row, New York.

[Hicks, 2019] Hicks, M. (2019). What everyone should say about symmetries (and how humeans get to say it). *Philosophy of Science*, 86:1284–1294.

[Khalfin and Tsirelson, 1992] Khalfin, L. A. and Tsirelson, B. S. (1992). Quantum/classical correspondence in the light of Bell's inequalities. *Foundations of Physics*, 22:879–948. https://www.tau.ac.il/ tsirel/download/quantcl.ps.

[Knight, 2008] Knight, R. (2008). *Physics for Scientists and Engineers with Modern Physics*. Pearson, San Francisco.

[Landau, 1987] Landau, L. J. (1987). On the violation of Bell's inequality in quantum theory. *Physics Letters A*, 120(2):54–56.

[Menon, 2019] Menon, T. (2019). Algebraic fields and the dynamical approach to physical geometry. *Philosophy of Science*, 86:1273–1283.

[Mermin, 1981] Mermin, N. (1981). Bringing home the atomic world: Quantum mysteries for anybody. *American Journal of Physics*, 49(10):940–943.

[Mermin, 2019] Mermin, N. D. (2019). Making better sense of quantum mechanics. *Reports on Progress in Physics*, 82(1):012002. https://arxiv.org/abs/1809.01639.

[Norton, 2008] Norton, J. (2008). Why constructive relativity fails. *British Journal for the Philosophy of Science*, 59(4):821–834.

[Salmon, 1993] Salmon, W. C. (1993). The value of scientific understanding. *Philosophica*, 51(1):9–19.

[Serway and Jewett, 2019] Serway, R. and Jewett, J. (2019). *Physics for Scientists and Engineers with Modern Physics*. Cengage, Boston.

[Silberstein and Stuckey, 2020] Silberstein, M. and Stuckey, W. (2020). Rethinking the world with neutral monism: Removing the boundaries between mind, matter, and spacetime. *Entropy*, 22(5):551. https://doi.org/10.3390/e22050551.

[Silberstein et al., 2018] Silberstein, M., Stuckey, W., and McDevitt, T. (2018). *Beyond the Dynamical Universe: Unifying Block Universe Physics and Time as Experienced*. Oxford University Press, Oxford, UK.

[Silberstein et al., 2021] Silberstein, M., Stuckey, W., and McDevitt, T. (2021). Beyond causal explanation: Einstein's principle not reichenbach's. *Entropy*, 23(1):114. https://www.mdpi.com/1099-4300/23/1/114/htm.

[Stuckey et al., 2019] Stuckey, W., Silberstein, M., McDevitt, T., and Kohler, I. (2019). Why the Tsirelson bound? Bub's question and Fuchs' desideratum. *Entropy*, 21:692. https://arxiv.org/abs/1807.09115.

[Stuckey et al., 2020] Stuckey, W., Silberstein, M., McDevitt, T., and Le, T. (2020). Answering Mermin's challenge with conservation per no preferred reference frame. *Scientific Reports*, 10:15771. www.nature.com/articles/s41598-020-72817-7.

[Van Camp, 2011] Van Camp, W. (2011). Principle theories, constructive theories, and explanation in modern physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 42:23–31.

[Weinberg, 2017] Weinberg, S. (2017). The trouble with quantum mechanics. https://www.nybooks.com/articles/2017/01/19/trouble-with-quantum-mechanics/.

[Wheeler, 1986] Wheeler, J. (1986). How come the quantum? *New Techniques and Ideas in Quantum Measurement Theory*, 480:304–316. https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1986.tb12434.x.

To be presented at the 27th biennial Philosophy of Science Association meeting.

Epistemic Risk in the Triangulation Argument for Implicit Attitudes

Morgan Thompson

One important strategy for dealing with error in our methods is triangulation, or the use multiple methods to investigate the same hypothesis. Current accounts of triangulation focus the conditions under which it succeeds, but ignore the many ways it can fail in practice. Instead, I argue that an account of triangulation focused on epistemic risk is better able to describe how triangulation fails and to normatively guide future triangulation research.

In this paper, I defend the claim that a useful account of methodological triangulation needs to account for the ways triangulation is susceptible to failure in its practice rather than focusing primarily on how and why it succeeds in ideal cases. A theory or account of a practice should highlight potential failures in order to be useful. Consider some ethical theory that gives an account of right and wrong actions. In order to use this ethical theory to guide my actions, I need to know not just what makes an action right or wrong, but also some features of my moral psychology. What are the ways that I am likely to err? Should I be worried about having a weak will and lacking follow through for actions that I deem right? Knowledge of the ways in which I might err allows me to better use the ethical theory to guide my actions. Analogously, I argue that an account of triangulation that is useful in practice ought to explain not just why triangulation is successful in ideal cases, but also how it can fail in practice. To do so, I will appeal to the idea of epistemic risk from the literature on the types and roles of values in science, medicine, and technology. By identifying types of failure, this lays the groundwork for future normative work developing strategies to avoid or mitigate these risks in triangulation research.

To be presented at the 27th biennial Philosophy of Science Association meeting.

1.1 Methodological Triangulation

Methodological triangulation involves the use of multiple methods to examine the same research question. Current accounts of triangulation are cashed out in terms of its success.[1] One view of triangulation sets out to: "identify at an abstract level the logic behind successful robustness arguments [and…] to determine what is required for a specific form of robustness analysis to be successful" (Kuorikoski and Marchionni 2016, 230). On another view, triangulation is defined as: "the use in empirical practice of multiple means of investigation to validate an experimental outcome" (Schickore and Coko 2013, 296). Current accounts agree on two success criteria: (i) the methods employed need to be sufficient diverse and (ii) the methods need to produce data about the same phenomenon.

How would this received view of triangulation account for cases of failure in practice? There is substantial discussion of the failure to have sufficiently diverse methods (i), which is what Wimsatt (1981) called "illusory robustness." Still these accounts of diversity are based on successful cases of triangulation (e.g., Schupbach 2018).

We can also consider the other success criterion in triangulation: that each method produces data about the same phenomenon (ii). While most philosophers working on triangulation recognize that this is a success criterion, relatively little has been said about how researchers can *know* they

[1] One exception is Stegenga (2009) who considers various problems with the use of triangulation as a strategy to deal with the problem of epistemic uncertainty in science. However, many of his critiques are not internal to the practice of triangulation. Stegenga's main concern is that philosophical accounts of triangulation provide no guidance when evidence both confirms and disconfirms the same hypothesis. But most centrally to this paper, Stegenga does not examine the epistemic risks triangulation arguments are subject to when they *appear* to be successful. These potential errors are all the more suspect because they mascaraed as successes.

To be presented at the 27th biennial Philosophy of Science Association meeting.

have met this criterion.[2] Even less has been said about how researchers can fail to meet this success criterion.

1.1.1 Epistemic Risk

In order to flesh out an account of triangulation that explains how it can fail in practice, I appeal to the concept of epistemic risk, which is "any risk of epistemic error that arises anywhere during knowledge practices" (Biddle and Kukla 2017, 218). There are many types of epistemic risk that occur at different parts of the research process. The most discussed kind of epistemic risk is inductive risk (Douglas 2016), which is particularly predominant in discussion about the role of values in science, medicine, and technology. Although the name implies it is any risk in inductive inferences, it is a technical term that refers specifically to the risk in inductive inferences from evidence to acceptance or rejection of a hypothesis.

Following Biddle & Kukla (2017), I hold that focusing exclusively on inductive risk makes our philosophical accounts of epistemic risk deficient. Other types of epistemic risk include the risk in deciding whether to characterize some datum as evidence for a hypothesis, such as whether some particular slide contains tumors and whether the tumors were malignant (Biddle's (2016) interpretation of Douglas 2000, 569). Another example is risk in the inference from animal models to the target system of interest (usually in humans) as in research on exposure to bisphenol A in a particular rat model (Biddle's (2016) interpretation of Wilholt 2009).

[2] One exception is Kuorikoski and Marchionni (2016), who argue that triangulation primarily consists in justifying data-to-phenomena inferences. Relying on Bogen and Woodward (1988), Kuorikoski and Marchionni argue that researchers can use empirical reasoning to justify these inferences, such as intervening on the phenomenon to determine whether there are corresponding differences in the data. While I think their view is on the right track, it is (1) susceptible to the criticism of not explaining why triangulation sometimes fails and (2) does not provide a sufficiently developed account of the practice of triangulation. I aim to rectify these two issues here.

To be presented at the 27th biennial Philosophy of Science Association meeting.

Current accounts of triangulation focused on success can only account for two types of epistemic risk: the failure to have sufficiently diverse methods (or Wimsatt's "illusory robustness") and, on my view, inductive risk. I will argue that an account of triangulation that explains failure will need to make use of epistemic risk more broadly as not all instances fall neatly under the risk of illusory robustness or inductive risk.

1.1.2 Schema for Triangulation in Practice

In order to develop an account of triangulation that highlights points of failure, I turn away from abstract success conditions and to the details of knowledge production via triangulation. I highlight important steps in the practice of triangulation from the causal production of data to its transition to playing an evidential role to the increased credence in some hypothesis. In this section I provide a schema for the practice of triangulation.

Let me first distinguish between data and phenomena (Bogen and Woodward 1988). Data are publicly observable reports that result from experimental or observational processes. They are not repeatable because they are the actual reports produced through experimentation or observation. Phenomena on the other hand are stable patterns in the world. Phenomena are often not directly observable and are characterized and explained by theory.

In the practice of triangulation, researchers identify multiple methods that are likely to produce data relevant to the same phenomenon. Each method may include some sources of error, such as random error from sampling or systematic error due to the instruments and procedures of the method. Unfortunately, researchers are often unaware of all sources of error in their methods. And these errors causally impact what data is produced. Yet, it is this data produced by imperfect methods that is the input for our inferential reasoning.

To be presented at the 27th biennial Philosophy of Science Association meeting.

Here let me make a further distinction between data and evidence. Rather than thinking of evidence as a separate kind of entity, we can think of it as a role that data play in confirming or disconfirming some hypothesis. In some cases of triangulation, this step may not be trivial: when data is produced in radically different experimental and theoretical contexts, many assumptions may be required to get from these different datasets to evidence that bears on (some particular) hypothesis. This problem about the evidential role of data is what Stegenga (2009) calls this the problem of incongruity.

Consider also that the data may be used as evidence in relation to multiple hypotheses. That is, despite of the fact that it may have been collected with some particular purpose in mind, it can serve as evidence for or against other hypotheses. In the case of triangulation, we're interested only in data that can be used as evidence for the same hypothesis. I'll focus on hypotheses about the existence of a phenomenon, though triangulation can also be used to estimate parameters and constants (e.g., Avogadro's number). At this point in the practice of triangulation, it needs to be demonstrated that all of the diverse datasets can serve as evidence for or against the *same* hypothesis.

Then once the evidential role of the datasets with respect to the same hypothesis has been established, researchers can make an inference to accept or reject the hypothesis. Even if all of the datasets provide supporting evidence for the hypothesis, a judgement still needs to be made about whether sufficient evidence has been collected to accept the hypothesis.

Theory can help reduce the uncertainty for some cases of triangulation. If researchers are triangulating on a claim about the existence of a phenomenon, then they should use some theoretical characterization of that phenomenon that describes its features. Researchers need a

To be presented at the 27th biennial Philosophy of Science Association meeting.

sufficiently developed characterization of a phenomenon in order to distinguish between inferences to the phenomenon of interest from inferences to other phenomena.

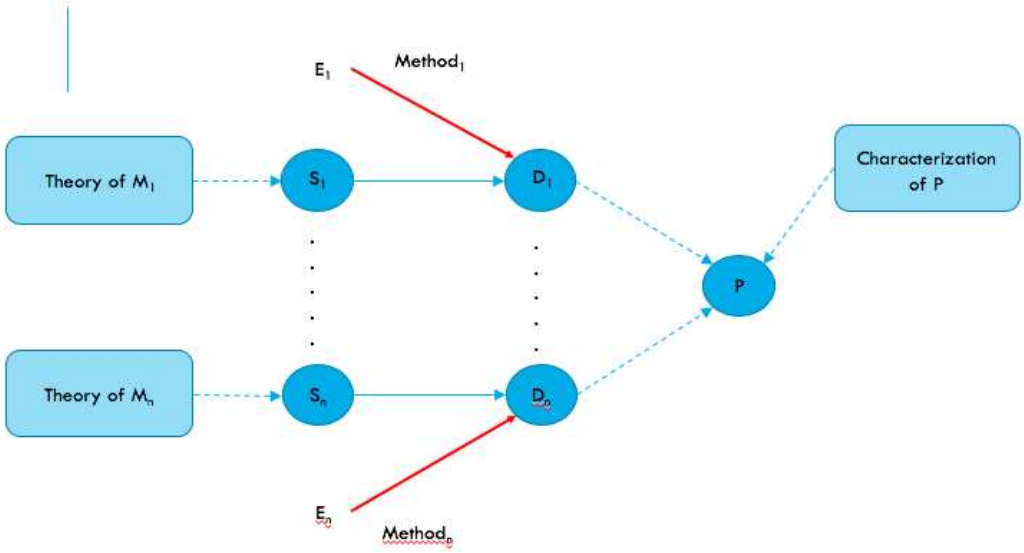


Figure 1. Schema of Triangulation

1.2 Triangulation in Implicit Social Cognition

Now that I've described the process of triangulation, I will demonstrate how it locates different types of epistemic risk. To do so, I will analyze the triangulation argument for implicit attitudes in social psychology.

By the mid-1990s, the majority of participants in psychology studies no longer self-reported holding explicitly racial attitudes (e.g., Dovidio and Gaertner 2000). In fact, many participants began to view racist acts as socially unacceptable and avoided committing racist actions themselves (Sue 2010). Yet, widespread racially discriminatory practices and racial

To be presented at the 27th biennial Philosophy of Science Association meeting.

disparities in economic, social, and health spheres persisted. Social psychologists posited that an explanation for these apparently contradictory features was that individuals still held racially biased attitudes, but that they were not reporting them when asked directly about their attitudes. So, researchers developed new techniques to control for the social desirability of appearing egalitarian (e.g., the "bogus pipeline" Jones and Sigall 1971). Indirect measures get around participants' ability and motivation to present themselves in a particular way to the researchers and instead measure their less controlled responses. As a result, researchers posited 'implicit attitudes' as a mental state or process. Implicit attitudes are automatically activated evaluative judgments about which participants are typically unaware or unable to control.

1.2.1 The IAT and the Evaluative Priming Task

The study of implicit attitudes bloomed. There are now nearly two dozen methods for measuring implicit attitudes. The two initial and most well-developed of these methods are the Implicit Association Test (IAT) (e.g., Greenwald, McGee, and Schwartz 1998) and the evaluative priming task (EPT) (e.g., Fazio et al. 1986). I discuss each in turn.

During a racial IAT, participants view stimuli from four categories: two racial groups and two evaluative groups. On any trial, each racial group is paired with a different evaluative category and these pairing are displayed on either side of the display screen. On typical racial IATs, two of the categories are stimuli related to two racial groups (e.g., faces of White and Black individuals) and two of the categories are evaluative stimuli (e.g., positive and negative words). Participants are asked to quickly categorize stimuli by pressing one of two keys on the corresponding to the disjunctive categories listed on the right and left sides of the display. Researchers can compare participants' reaction times on trials in which Black-positive and White-negative are paired to

those in which Black-negative and White-positive are paired. A faster response time to the latter compared to the former is thought to indicate racial attitudes that more closely link Black people with negative concepts and White people with positive concepts (e.g., Mitchell, Nosek, and Banaji 2003).

Evaluative priming tasks instead use stimuli from the categories of interest to prime participants before participants perform a categorization task on unrelated evaluative target stimuli. If researchers are interested in racial attitudes, they might use images of Black or White people to prime participants. Then during the categorization task, participants are asked to categorize positive- and negative-valence words (target stimulus). Researchers reason that reaction times on the categorization task will be influenced by the evaluative valence of the prime stimulus. If a participant holds negative attitudes towards White people, then after viewing a White stimulus prime, they will categorize negative target words more quickly than positive target words.

1.2.2 The Triangulation Argument for Implicit Attitudes

Social psychologists take indirect measures like the IAT and EPT to triangulate on the same phenomenon—implicit attitudes. Over time, theories about how to characterize implicit attitudes have changed, but the assumption that the triangulation argument for implicit attitudes is successful has remained. Here I will offer some evidence for this claim.

Discussing the views of the field at the time in a review article on the nature of implicit attitudes, Gawronski, Hofmann, and Wilber (2006, 486; citations removed) state:

> A widespread assumption underlying the application of indirect measures is that they provide access to unconscious mental associations that are difficult to assess with standard self-report measures. Specifically, it is often argued that self-reported (explicit) evaluations reflect conscious attitudes, whereas indirectly assessed (implicit) evaluations reflect unconscious attitudes.

To be presented at the 27th biennial Philosophy of Science Association meeting.

While Gawronski and colleagues go on to critique this widespread assumption (at least, its attribution of 'unconscious' to implicit attitudes), this quote demonstrates the ubiquitous assumption among implicit attitude researchers that first-generation indirect methods measured implicit attitudes.

More recently social psychologists have developed a neutral characterization of implicit attitudes that does not commit to any particular view of 'implicit'. This is to broadly accommodate issues that participants are able to predict the evaluative direction of their implicit attitudes (Hahn et al. 2014). As Greenwald and Lai write in a review article this year, "The currently dominant understanding of "implicit" among social cognition researchers is "indirectly measured." The labels "indirectly measured attitude" and "implicit attitude" are used interchangeably in this review" (Greenwald and Lai 2020). Still the assumption remains: whatever indirect measures are measuring, it is the same phenomenon.

1.3 Two Epistemic Risks in Triangulation

In this section, I use my account of triangulation to highlight two examples of epistemic risks and where they arise in implicit attitude research. My account better explains what goes wrong in these cases than accounts of triangulation focused on success. That is, my account provides a better descriptive account of scientific practice, where triangulation does not always succeed. Here I identify two types of epistemic risk: (1) epistemic risk when data is taken to be evidence for some hypothesis and (2) inductive risk in determining a sufficient level of evidence for the acceptance or rejection of a hypothesis.

To be presented at the 27th biennial Philosophy of Science Association meeting.

1.3.1 Moving from Data to Evidence

One major epistemic risk in triangulation is that we may mistakenly think that the different datasets can serve as evidence for the same hypothesis. We are particularly at risk of this error when we do not justify the claim that our methods measure aspects theoretically related to the same hypothesis. Data do not automatically bear on hypotheses. A datum can be an image from electron microscopy, a mark selecting an answer on a survey, or recorded video of a researcher interacting with participants. So, data needs to be interpreted in relation to the hypotheses for which they may serve as evidence. In doing this, researchers must infer on the basis of data and some assumptions to the confirmation or disconfirmation of a hypothesis.

I argue that this epistemic risk is relevant to the triangulation argument for implicit attitudes. The data produced and current assumptions in social psychology do not support the claim that the data produced by the IAT and EPT serve as evidence for the same hypothesis. In fact, according to some implicit attitude researchers, they serve as evidence for slightly different hypotheses.

In IAT studies, the categories of interest are made explicit to the participant as the categories must be identified and paired to perform the categorization task. Thus, IAT scores are thought to measure attitudes toward the general social category. Thus, they can serve as evidence for hypotheses about associations between evaluative categories and social categories.

In an evaluative priming task, on the other hand, the instructions do not explicitly determine the relevant categorical membership of the priming stimulus. It is generally accepted that due to this feature, evaluative priming tasks measure attitudes toward the stimuli rather than the category (Olson and Fazio 2003; Mitchell, Nosek, and Banaji 2003). Consider that the priming stimulus is often an image of a person's face. Researchers may wish to contrast Black

To be presented at the 27th biennial Philosophy of Science Association meeting.

and White faces as priming stimuli in an evaluative priming task; however, as a feature of the images individuals represented will also belong to other social categories (e.g., attractiveness, gender). Because the categorization task is only along the evaluative dimension, it is not made salient which of these categories a participant is responding to. Consider the case of a participant who when primed with a particular image of a Black face, categorizes positive stimuli more slowly than when primed with an image of a White face. The response discrepancy could be caused by a negative evaluations of the person-represented-in-the-image's perceived race, attractiveness, perceived gender, or any combination of these and other features.

Good task design will control for these differences as much as possible, but due to the design of the task, it is impossible to identify what features influence the participant's reaction times in the categorization task in any given case. The features that cause a response discrepancy may change over time even for the same participant because implicit attitudes are thought to be context dependent (Jost 2019) and the empirical findings that indirect measures generally have low test-retest validity (Bosson, Swann, and Pennebaker 2000).

In order to address this epistemic risk, researchers need to provide justification for the claim that the IAT and EPT produce data that can serve as evidence for the hypothesis that participants have a negative association with the social category of interest. For the IAT, this justification already exists. For the EPT, it is less obvious. So, using my account of triangulation, I have highlighted a particular weak point in the triangulation argument for implicit attitudes and emphasized a place for the development and elaboration of norms for successful triangulation. Note also that this epistemic risk does not fit neatly under the heading "illusory robustness" or inductive risk because the problem arises due to the differences in the methods and does not involve a judgement about accepting or rejecting a hypothesis.

To be presented at the 27th biennial Philosophy of Science Association meeting.

1.3.2 Inductive Risk

Once we know data can serve as evidence for the same hypothesis, we can ask: How do researchers know there is sufficient evidence to accept the hypothesis? On my view, the epistemic risk of error here is best characterized as inductive risk. However, in the context of triangulation inductive risk takes a particular form. Specifically, researchers ought to be concerned about the risk of accepting the hypothesis when it is false. In cases where our hypothesis is about the existence of some phenomenon (as triangulation is often used), the inductive risk may be specifically sensitive to the error that data produced (and their evidential support) are actually for distinct phenomena. In other words, there is an inductive risk in accepting the hypothesis that some phenomenon of interest exists on the basis of triangulation, especially when we have not sufficiently ruled out the possible hypothesis that multiple phenomena are differentially driving the results.

Psychologists evaluate the validity of their tests using psychometrics. Relevant to my arguments, convergent validity is the extent to which two methods that are predicted to measure the same phenomenon are in fact measuring the same phenomenon. Low convergent validity suggests that two methods measure different phenomena. Psychologists often assess convergent validity by examining correlation coefficients.[3] If two methods measure the same phenomenon, they are expected to have high correlations in their scores. However, given that the two methods are distinct in some ways, there should not be a perfect correlation in their scores. There is no

[3] Other methods such as the multi-trait multi-method matrix (Campbell and Fiske 1959) have been used less frequently and less completely in the context of implicit attitudes.

To be presented at the 27th biennial Philosophy of Science Association meeting.

well accepted threshold for what counts as sufficiently high convergent validity. But social psychologists hold that the IAT and EPT ought to have high convergent validity (e.g., Banaji 2001).

Unfortunately, researchers have found low correlations between the IAT and other implicit measures and thus, low convergent validity (Fazio and Olson 2003). The correlation in scores for the IAT and EPT range between r=.24 and r=.13. These are very low positive correlations. So, a participant's score on the IAT provides very little information about their EPT score, and vice versa.

One possible cause of the low correlations between IAT and EPT scores is the low reliability of EPT (De Houwer et al. 2009). Perhaps the scores do not correlate well due to noisiness in the data produced by unreliable methods rather than the methods measuring different phenomena. A recent comparison of seven indirect measures of attitudes Bar-Anan and Nosek (2014), the EPT had weak correlations with other indirect measures (including the IAT, r=.24).

However, there are two reasons to remain neutral with respect to these explanations. First, a measure need not be reliable for it to be valid (Borsboom, Mellenbergh, and Van Heerden 2004). The measure could track a context-dependent phenomenon, of which implicit attitudes is probably an example (Jost 2019). Second, as Bar-Anan and Nosek (2014, 677, original emphasis) suggest, low convergent validity and low reliability may *both* contribute to the low correlations of scores on indirect measures of attitudes:

> the most likely explanation for this pattern, coupled with the similar rank ordering for internal consistency, is that [Affective Misattribution Priming] and EPT are both relatively distinct, and *also* less effective in reliably assessing the target evaluation than are the other measures. […] it could still be the case that both measures assess unique components of evaluation that are not assessed by other indirect measures (including each other).

To be presented at the 27th biennial Philosophy of Science Association meeting.

Still one promising finding is that unlike the Affective Misattribution Priming task, Bar-Anan and Nosek (2014) do not find a strong correlation between the EPT and direct measures of racial attitudes (i.e., self-report on surveys), which would have indicated the potential influence of deliberate evaluation in the indirect measurement. So, while some of the low correlations between the measures may be due to the low reliability of the EPT, it is possible that both low reliability and low convergent validity are part of the picture.

1.3.3 Why can't these be understood as a failure of diversity?

One potential objection is that the IAT and EPT are not sufficiently diverse methods. The basic idea is that whatever diversity criterion we accept (see Schupbach 2018), the IAT and EPT are too similar to count as distinct methods for the purposes of triangulation. I respond to this objection by clarifying that these methods historically descendant from different theories in psychology. In addition to my arguments that they produce data relevant to different hypotheses (section 1.3.1), this gives us some reason to think the methods are sufficiently diverse on any appropriate diversity criterion.

The two methods I discuss were developed out of different historical traditions in psychology (Payne and Gawronski 2015). Drawing on Shiffrin and Schneider's (Shiffrin and Schneider 1977) work on selective attention, Fazio and colleagues (Fazio, Jackson, Dunton, & Williams, 1995) developed the evaluative priming task to distinguish automatic and controlled processing. Controlled processing requires attention and can be altered voluntarily, whereas automatic processing takes place on memories stored in long-term memory, is automatically activated given the appropriate inputs, and is difficult to suppress.

To be presented at the 27th biennial Philosophy of Science Association meeting.

Greenwald and Banaji's (1995) work on implicit attitudes came out of cognitive psychological research on implicit memory, which describes the way that earlier experiences can influence current performance on learned tasks without conscious awareness of the past experiences. Most famously, the patient H.M., who had a medial temporal lobectomy and thus lacked bilateral hippocampi and other structures, was unable to create new episodic memories. However, H.M. demonstrated the formation of new implicit memories through the time-savings in relearning motor skill tasks (Corkin 2002). As Greenwald et al. (1998) constructed it, the IAT is a measurement of implicit memory. So, both measures were designed based on different theories. In short, the evaluative priming task was designed to measure a construct that is typically uncontrolled or automatic while the IAT is designed to measure a construct that is typically unconscious or about which the individual is unaware.

1.4 Conclusion

In this paper, I have provided an account of triangulation that highlights locations and types of epistemic risk. In particular, I diagnosed two epistemic risks in implicit attitude research: (1) the risk that data do not serve as evidence for the same hypothesis, and (2) the particular inductive risk that there is insufficient evidence provided to conclude that there is a single phenomenon (given the plausibility of alternative hypotheses positing multiple phenomena). Neither is sufficiently described by illusory robustness and (1) is not a case of inductive risk either. Finally, I demonstrated that current accounts of triangulation focused on successful cases cannot provide explanations of why triangulation sometimes fails in practice and thus, do not develop sufficient norms to guide future triangulation research.

To be presented at the 27th biennial Philosophy of Science Association meeting.

References

Banaji, Mahzarin R. 2001. "Implicit Attitudes Can Be Measured." In *The Nature of Remembering: Essays in Honor of Robert G. Crowder*, edited by H.L. Roediger and J.S. Nairne, 117–50. Washington, DC: American Psychological Association.

Bar-Anan, Yoav, and Brian A. Nosek. 2014. "A Comparative Investigation of Seven Indirect Attitude Measures." *Behavior Research Methods* 46(3):668–88.

Biddle, Justin B., and Rebecca Kukla. 2017. "The Geography of Epistemic Risk." *Exploring Inductive Risk: Case Studies of Values in Science*, 215–38.

Biddle, Justin B. 2016. "Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease." *Perspectives on Science* 22(3):397–417.

Bogen, James, and James F Woodward. 1988. "Saving the Phenomena." *The Philosophical Review* XCVII(3):303–52.

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111(4):1061–71.

Bosson, J K, W B Swann, and J W Pennebaker. 2000. "Stalking the Perfect Measure of Implicit Self-Esteem: The Blind Men and the Elephant Revisited?" *Journal of Personality and Social Psychology* 79:631–643.

Campbell, Donald T, and Donald W Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56(2):81–105.

Corkin, Suzanne. 2002. "What's New with the Amnesic Patient H.M.?" *Nature Reviews Neuroscience* 3(2):153–60.

Douglas, Heather. 2016. "Values in Science." *Oxford Handbook in The Philosophy of Science*, 23.

Dovidio, John F, and Samuel L Gaertner. 2000. "Aversive Racism and Selection Decisions: 1989 and 1999." *Psychological Science* 11(4):315–19.

Fazio, Russell H., Joni R. Jackson, Bridget C. Dunton, and Carol J. Williams. 1995. "Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?" *Journal of Personality and Social Psychology* 69(6):1013–27.

Fazio, Russell H., and Michael A. Olson. 2003. "Attitudes: Foundations, Functions, and Consequences." *The SAGE Handbook of Social Psychology*, no. January:123–45.

Fazio, Russell H, David M Sanbonmatsu, Martha C Powell, and Frank R Kardes. 1986. "On the Automatic Activation of Attitudes." *Journal of Personality and Social Psychology* 50(2):229–38.

Gawronski, Bertram, Wilhelm Hofmann, and Christopher J. Wilbur. 2006. "Are 'Implicit' Attitudes Unconscious?" *Consciousness and Cognition* 15:485–99.

Greenwald, Anthony G., and Mahzarin R Banaji. 1995. "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes." *Psychological Review*.

Greenwald, Anthony G., D E McGee, and J L K Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74:1464–1480.

Hahn, Adam, Charles M. Judd, Holen K. Hirsh, and Irene V. Blair. 2014. "Awareness of Implicit Attitudes." *Journal of Experimental Psychology: General* 143(3):1369–92.

Houwer, Jan De, Sarah Teige-Mocigemba, Adriaan Spruyt, and Agnes Moors. 2009. "Implicit Measures: A Normative Analysis and Review." *Psychological Bulletin* 135(3):347–68.

Jones, Edward E., and Harold Sigall. 1971. "The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude." *Psychological Bulletin* 76(5):349–64.

Jost, John T. 2019. "The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit

To be presented at the 27th biennial Philosophy of Science Association meeting.

Attitudes Are Indispensable to Social and Political Psychology." *Current Directions in Psychological Science* 28(1):10–19.

Kuorikoski, Jaakko, and Caterina Marchionni. 2016. "Evidential Diversity and the Triangulation of Phenomena." *Philosophy of Science* 83(2):227–47.

Mitchell, J P, Brian A. Nosek, and Mahzarin R. Banaji. 2003. "Contextual Variations in Implicit Evaluation." *Journal of Experimental Psychology: General* 132:455–469.

Olson, Michael A, and Russell H Fazio. 2003. "Relations between Implicit Measures of Prejudice: What Are We Measuring?" *Psychological Science* 14:636– 639.

Payne, B Keith, and B Gawronski. 2015. "A History of Implicit Social Cognition: Where Is It Coming From? Where Is It Now? Where Is It Going?" In *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, edited by B. Gawronski and B K Payne. New York: Guilford.

Schickore, Jutta, and Klodian Coko. 2013. "Using Multiple Means of Determination." *International Studies in the Philosophy of Science* 27(3):295–313.

Schupbach, Jonah N. 2018. "Robustness Analysis as Explanatory Reasoning." *British Journal for the Philosophy of Science* 69(1):275–300.

Shiffrin, Richard M., and Walter Schneider. 1977. "Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending and a General Theory." *Psychological Review* 84(2):127–90.

Stegenga, Jacob. 2009. "Robustness, Discordance, and Relevance." *Philosophy of Science* 76:650–61.

Wilholt, Torsten. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science Part A* 40(1):92–101.

The Epistemic Value of the Living Fossils Concept

March 6, 2020

Word Count: 4,985

Abstract

Living fossils, taxa with similar members now and in the deep past, have recently come under scrutiny. Those who think the concept should be retained have argued for its epistemic and normative utility. This paper extends the epistemic utility of the living fossils concept to include ways in which a taxon's living fossil status can serve as evidence for other claims about that taxon. I will use some insights from developmental biology to refine these claims. Insofar as these considerations demonstrate the epistemic utility of the living fossils concept, they support retaining the concept and using it in biological research.

Living fossils are taxa in which extant organisms morphologically resemble fossilized organisms; paradigmatic examples include horseshoe crabs, coelacanths, and tuataras. Recently the living fossil concept has received considerable criticism, with even paradigmatic cases being contested. Some argue that the concept is not very useful for biologists, since these diverse cases are unlikely to be the product of unified phenomena, while others argue that the concept may be useful for certain epistemic and normative

purposes. My aim in this paper is to address the epistemic value of living fossils. In particular, I will address the following question: Given that a taxon is a living fossil, what else do we know about it? Using considerations from developmental biology, I show that many common inferences from morphological similarity fail in the context of living fossils. I will argue, however, that there are some inferences that are justified. I conclude that the living fossil concept has epistemic value, and hence should be retained.

After reviewing the recent literature (section 1), I will address three obvious conclusions that we might want to draw about living fossils (section 2): (1) non-morphological phenotypic similarity between the extant and past taxa, (2) the existence of a persistent lineage that includes these taxa, and (3) a slow rate of evolutionary change between these taxa. I will evaluate each of these inferences, especially using insights from developmental biology (section 3).

1 Defining 'living fossil'

Philosophers of biology have offered different characterizations of living fossils. Lidgard and Love (2018) argue for ways in which the concept is useful in setting research agendas, despite ambiguity in whether particular taxa should be classified as living fossils. Turner (2019) suggests an explicit definition of living fossil, one which he believes enables us to use living fossils to set conservation priorities. Specifically, Turner thinks that living fossils are taxa which have:

1. Prehistorically deep morphological stability,
2. Few extant species, and
3. High contribution to phylogenetic diversity.

Werth and Shear (2014) give a similar characterization of living fossils, picking out "morphological conservatism" and "little taxonomic diversity" as relevant factors (434, 436).

Turner (2019) thinks there is epistemic value to the living fossil concept, including that "observations of [extant organisms in a living fossil taxon] can surely tell us something about the prehistoric ones" (11). The next two sections of this paper will specify exactly what we might be able to learn about these prehistoric taxa on the basis of their living fossil status. To sidestep debates about the specific definition of "living fossil," I will focus on the epistemic role of morphological similarity between past and extant taxa, a feature unanimously associated with living fossils.

Note that this paper is concerned with the possibility that the living fossil concept is *epistemically* valuable, although it may be valuable in other ways, including for normative purposes (as Turner 2019 argues). One way in which the living fossils concept might be epistemically valuable is that it helps us identify evolutionary episodes in need of explanation. Lidgard and Love (2018) think this is one purpose of the concept. In this case, a taxon's living fossil status, or at least the various features associated with that status, is the *explanandum*. However, in the remainder of this paper, I focus on another possible epistemic role for the living fossils concept to serve: a taxon's living fossil status can serve as *evidence* for other claims about the members of that taxon.

2 Inferences from morphological similarity

To reject the arguments of skeptics who think we should do away with the living fossil concept (e.g., Casane and Laurenti 2013, Mathers et al. 2013, Wagner et al. 2017), we

should show what role the concept can play. Lidgard and Love and Turner recognize this, although they have different ideas of what this role is. However, the authors seem to be in agreement that part of what we want to be able to use the living fossil concept for is making inferences from the fact that past and extant taxa are morphologically similar to some other fact F about these taxa. For short:

$$\textit{morphological similarity} \rightarrow F \tag{1}$$

Both Lidgard and Love and Turner agree that we should be able to use the living fossil concept to make inferences of this form. Turner (2019) calls this the "epistemic value" of focusing on the morphological resemblance of past to extant taxa (11).

One possible fact F that we may want to infer from morphological similarity between two taxa is that these taxa are phenotypically similar in ways above and beyond their morphological similarity. Take horseshoe crabs. Extant horseshoe crabs have hemocyanin in their blood (they use copper rather than iron to transport oxygen). Turner (2019) says, "the fossil record does not tell us that ancient horseshoe crabs had hemocyanin in their blood. But that seems like a fairly safe inference, given our background knowledge of phylogeny plus the observation that living ones do have hemocyanin in their blood" (11). The general type of inference that Turner is making is something like:

$$\textit{morphological similarity} \rightarrow \textit{general phenotypic similarity} \tag{2}$$

So far, I have been talking about morphological *similarity*, rather than morphological *stability*, the latter of which is used in Turner's definition. Turner (2019) says that showing morphological stability between past and extant taxa is equivalent to showing

morphological similarity within a persistent lineage (3). If morphological similarity itself was evidence for persistence of a lineage, then morphological similarity would be evidence for morphological stability. In other words, the following inferences are equivalent:

$$morphologial\ similarity \rightarrow persistence\ of\ lineage \tag{3}$$

$$morphological\ similarity \rightarrow morphological\ stability \tag{4}$$

Finally, the living fossil concept may be useful for inferring rates of evolutionary change:

$$morphological\ similarity \rightarrow slow\ evolutionary\ rate \tag{5}$$

It only makes sense to talk about a rate of evolution *within* a given lineage, so inference 3 is necessary for inference 5.

Inferences 2, 3, and 5 do not exhaust the possible inferences from morphological similarity to F which we might make about living fossil taxa, but these examples show the possibility of making inferences about living fossil taxa based on what else we know about them. Thus these inferences provide good candidates if we want to demonstrate the epistemic utility of the living fossil concept.

The following section will use some insights from developmental biology to evaluate these inferences.

3 Developmental considerations

Various concepts and theories in evolutionary biology have been revised in light of results in developmental biology. For example, developmental plasticity provides a possible explanation of speciation events, one compatible with the theory of punctuated equilibrium (West-Eberhard 2003). On the basis of such results, some have even suggested replacing the Modern Synthesis with the Extended Evolutionary Synthesis (e.g., Laland et al. 2015).

In this section, I will use some results in developmental biology to examine the arguments using morphological similarity from section 2.

3.1 Non-morphological phenotypic similarity

Inference 2 says that we can infer from their morphological similarities that past and extant taxa have phenotypic similarities above and beyond these morphological similarities. For example, we would be able to infer the presence of hemocyanin in past horseshoe crabs on the basis that they are morphologically similar to extant horseshoe crabs.

Although I did not say this in section 2, one might have thought that the argument relating morphological similarity to general phenotypic similarity was implicitly assuming some relationship between morphological similarity and *genetic* similarity. If morphological similarity was good evidence for genetic similarity, and genetic similarity was good evidence for otherwise phenotypic similarity, then morphological similarity would be good evidence for phenotypic similarity. Including the implicit step, inference 2 would become:

morphological similarity → *genetic similarity* → *general phenotypic similarity* (6)

Even using a very rudimentary understanding of genetics, it is unlikely that inference 6 will work. The problem is that morphological similarity does not imply genetic similarity. There is not, in general, a one-to-one correspondence between genes and phenotypes, including morphology, so we can neither infer genetic information from phenotypic information nor vice versa. The same genes can result in different phenotypes, and the same phenotypes can be the result of different genes (e.g., Fusco and Minelli 2010). The former is the result of developmental plasticity, whereby a variety of environmental factors can affect phenotypic outcomes, for example by changing gene expression. The latter can be explained by the interchangeability of genes and environment in producing phenotypes,which West-Eberhard (2003) says "conflict[s] with the habit of supposing that the specificity of the [developmental] response comes entirely from the specificity of the gene" (117). If the argument from morphological similarity to general phenotypic similarity depends on an inference from morphological similarity to genetic similarity, the argument will fail, because the first half of inference 6 will turn out to be false. Additionally, and perhaps more intuitively, whatever genetic similarity might be implied by morphological similarity does not in itself imply the *additional* genetic similarity required to generate phenotypic similarity above and beyond morphology. In the case of the horseshoe crabs, different genes will be associated with morphology than with presence of hemocyanin.

However, morphological similarity may imply otherwise phenotypic similarity more

directly, as indicated in the original inference 2. For instance, we might think that certain non-morphological phenotypes are strongly correlated with particular morphologies.

Whether this correlation is plausible is going to depend on the non-morphological phenotype. For instance, whether extant horseshoe crabs' blood is similar to past horseshoe crabs' will depend on whether features of blood are strongly correlated with morphology. If we had independent evidence that the contents of blood and an organism's morphology were strongly correlated, then the inference from the horseshoe crabs' morphology to their blood phenotype would be unproblematic. However, as Lidgard and Love (2018) say, "[r]etention of some phenotypic (traditionally morphological) characters does not adequately *explain* change or the lack thereof in other phenotypic characters" (766, emphasis in original). Fortey (2011) also thinks that there can be "no final proof one way or the other" about whether the past horseshoe crabs' blood contained hemocyanin (27).

In fact, developmental biologists have recently stressed the *modularity* of phenotypes. This refers to the separability of phenotypes, despite possible integration among them; developmental modules are *semi-independent* and *dissociable*, meaning that various traits can occur in different combinations in different organisms, with varying degrees of interdependence between different traits (West-Eberhard 2003, chpt. 4). These modules can then be selected for separately. For instance, terrestrial and arboreal salamanders have distinct foot morphology; the developmental pathways that lead to these differences are relatively independent from the salamanders' other traits, which go (more or less) unaffected (Gilbert 2000). This is made possible by the branching nature of development: cell differentiation occurs at branching decision points, which can be triggered by genetic or environmental switches. West-Eberhard (2019) says that

modularity is a "universal property of organismic traits," because this branching process is ubiquitous (357). In the context of living fossils, and specifically inference 2, modularity means that morphological similarity – which may be dissociable or independent from other phenotypes – does not provide adequate evidence for similarity of non-morphological traits.

Of course, some modules are more interdependent, and can be expected to co-occur. For example, morphology can constrain behavior such that particular behavioral traits are strongly correlated with particular morphological traits. Whether presence of one phenotype provides good evidence for presence of another phenotype depends on having independent evidence of the ways in which the different developmental modules may be interdependent.

Therefore, the wholesale inference from morphological similarity to phenotypic similarity above and beyond morphology is unlikely to be justified. This is not just a general skepticism about our ability to infer the presence of some traits from the observation of others; developmental modularity gives us good reason to believe that many traits are dissociable. More specific cases, where a correlation between morphology and other phenotypes is independently established, may allow for appropriate use of this inference in living fossil taxa. Indeed, Lidgard and Love (2018) suggest that one of the questions that research on living fossils might be able to answer has to do with the role of developmental modularity in patterns of evolutionary stasis (766). In other words, we may be able to come to a better understanding of the ways in which different traits are combined in developmental modules by studying stasis of these traits in living fossil taxa.

3.2 Persistence of lineage

Inference 3 concludes on the basis of morphological similarity that the taxa are phylogenetically related such that they are both part of the same, persistent lineage, or, equivalently, that they are morphologically stable.

Note that neither Lidgard and Love nor Turner make the "lineage" relationship precise. Being part of the same lineage cannot require that the past fossil is an ancestor of the extant organisms, exactly, because we want to permit the past taxon and the extant one having an as-yet-unidentified common ancestor.[1] Neither can the lineage relationship be as broad as a whole clade; it would become meaningless to differentiate living fossils from other cases of relatedness between past and extant taxa. Although it is beyond the scope of this paper to more precisely say what a lineage is, I take it that it is something between an ancestor-descendant relationship and a clade.

Setting this aside: does morphological similarity imply persistence of lineage or morphological stability?

As in the case of phenotypic similarity, perhaps there is an implicit assumption contained in inference 3 that involves a relationship between morphology and genetics. Inference 3 could be justified on the basis of this relationship: if morphological similarity implies genetic similarity, and genetic similarity implies the phylogenetic relationship that would hold within a persistent lineage, then morphological similarity would imply persistence of lineage. The resulting inference is:

[1]This is likely the case with horseshoe crabs – see Fortey (2011).

$$morphological\ similarity \rightarrow genetic\ similiarity \rightarrow persistence\ of\ lineage \quad (7)$$

I have already argued in section 3.1 that morphological similarity does not imply genetic similarity, so inference 7 will not work.

However, we should consider whether morphological similarity implies persistence of lineage *without* relying on a connection to genetic similarity. I will argue that there are several reasons to think that it does not; however, morphological similarity is often the best evidence we have of phylogenetic relationships.

First, morphological similarity and persistence of lineage do not exactly imply morphological stability, because there is the possibility that the morphological trait was lost and reemerged within the same lineage. Alternatively, if the past and extant taxa are in the same clade but do not have an ancestor-descendant relationship, then it may be possible that their common ancestor was *not* morphologically similar, in which case the morphology would have had to emerge separately on two different branches of the phylogenetic tree. This would be a case of convergent evolution, where the same traits evolve twice. These considerations when checking for morphological stability are the same as the well-known issues with testing for homology (similarity due to common ancestry) in general.

Developmental biologists point out that developmental pathways, even if not morphological traits, may be homologous (e.g., Nijhout 2019, 946). In these cases, which are called parallelism (rather than convergence), the trait may appear to evolve separately in two different branches, or may appear to be lost and reemerge, when in fact

the mechanism by which the trait develops is actually homologous. It makes sense to broaden our concept of homology to include parallel evolution and recurrence of traits (West-Eberhard 2003, chpt. 25). I therefore concur with Turner (2019), although he does not explicitly use these developmental considerations to argue that morphological stability follows from morphological similarity.

Second, though, *lack* of morphological similarity may not be an indication of lack of morphological stability; polymorphism within a single species is relatively common. Two sample organisms from a species with morphologically distinct life stages may be mistaken as organisms belonging to different species if the organisms are observed in different of these life stages.[2] Extreme cases of sexual dimorphism are also liable to being mistaken for cases of multiple species. Note that both metamorphosis-induced life stages and sexual dimorphism may be the result of the developmental modularity discussed above (West-Eberhard 2003, 58, 75).

There is thus a risk of both false positives and false negatives in identifying persistence of lineage if we focus on morphological similarity. If there were a better indication of phylogenetic relationships than morphological similarity, we would use it instead.

These considerations notwithstanding, morphological similarity is often the best evidence we have for persistence of the same morphology over time, given that in the context of fossils we only have sporadic sample organisms and not any direct evidence of change over time.[3] This is part of the explanation for why the morphological species

[2]Turner (2016) acknowledges this point explicitly (64). See also Currie 2016.

[3]Note that our ability to acquire genetic information about fossil specimens may improve our epistemic position regarding phylogenetic relationships, if one thinks that the

concept – rejected nearly unanimously as an adequate species concept for extant species – is still used by paleontologists (e.g., Turner 2011, 49-50 and Werth and Shear 2014, 442-43). Often the best evidence we have for phylogenetic relationships involving fossils is morphological similarities and differences, and persistence of lineage in the context of living fossils is no different.

3.3 Evolutionary rates

The third candidate inference we might want to make from morphological similarity within living fossil taxa is a slow rate of evolutionary change between the past and extant taxa. Recall that the argument for a slow rate of evolutionary change requires that we accept the inference to persistence of lineage. I have suggested that morphological similarity is often the best evidence we can hope to have for persistence of lineage. In this section I will assume that that inference is justified, and move on to examining inference 5, from morphological similarity to a slow rate of evolutionary change.

As in sections 3.1 and 3.2, there is possibly an implicit assumption utilized here involving genetics. Let's ignore the possibility that the inference looks like this:

$$\textit{morphological similarity} \rightarrow \textit{genetic similarity} \rightarrow \textit{morphological stability} \rightarrow \textit{slow evolutionary rate} \tag{8}$$

because we are assuming that morphological similarity is directly evidence for morphological stability (and I have already argued that morphological similarity does

inference from genetic similarity to persistence of lineage is better than the inference from morphological similarity. See Jablonski and Shubin (2015).

not imply genetic similarity). In this case, the implicit justification for 5 is instead that morphological stability implies genetic stability, which in turn implies a slow rate of evolutionary change:

$$\textit{morphological similarity} \rightarrow \textit{morphological stability} \rightarrow \textit{genetic stability} \rightarrow \textit{slow evolutionary rate} \tag{9}$$

Many of the arguments I have given already that morphological similarity will not imply genetic similarity will be arguments against thinking that morphological *stability* implies genetic *stability*. I will not rehearse these arguments, because there is further reason to think that morphological stability does not imply genetic stability. Stabilizing selection acting on plastic traits can maintain the same phenotype over time, without necessarily having any effect whatsoever on rates of genetic change. For example, developmental plasticity is expected, especially in cases of extremely plastic traits like learning, to slow any directional increase or decrease in the propensity of a given phenotype in a population, because there is not ample opportunity for selection to act on any single phenotype (West-Eberhard 2003, 178). Furthermore, a process called "phenotypic accommodation" allows organisms to maintain functional phenotypic traits *despite* genetic mutation (West-Eberhard 2003, 51; see also West-Eberhard 2005).

The last step of inference 9 – from genetic stability to slow rate of evolutionary change – is also problematic, although my critique here will be more controversial. An intuitive view is that a slow rate of evolutionary change in a lineage *just is* a slow rate of genetic change in that lineage, and that therefore the move from genetic stability to slow rate of evolution is unproblematic (e.g., Schopff 1984, Ho 2008).

But is this really what we mean by slow rates of evolutionary change? Cases of stabilizing selection acting on phenotypes *without* causing a reduction in rates of genetic change show that it does not make sense to equate evolutionary change with genetic change. Traits on which stabilizing selection is acting should also be those traits which we say have a slow rate of evolutionary change: "the rate and degree of modification of a complex trait should be some positive function of its frequency of expression or use" (West-Eberhard 2003, 169). Traits with stability in a given lineage are exactly the traits with a slow rate of change. Therefore, there is no need to appeal to genetic stability to make the case for slow rates for evolutionary change – we can infer slow rates of evolutionary change directly from morphological stability.

Note that it is traits, and not lineages or taxa, to which we apply an evolutionary rate. Selection acts on phenotypes, not on organisms, species, or lineages. Lidgard and Love (2018) agree: "[c]haracters or character states are relatively more ancestral or derived, not whole organisms or lineages" (761, citing Omland, Cook, and Crisp 2008). Additionally, attribution of rates of change to traits rather than lineages is consistent with the idea of developmental modularity.

One of Turner's examples suggests that he thinks, in agreement with me, that morphological stability, the first feature in his definition of living fossils, is a better indication than molecular stability of slow rates of evolutionary change. Tuataras, a reptile from New Zealand, were thought to be living fossils on the basis of morphological stability, until researchers discovered that tuataras actually have a higher than average rate of molecular evolution (Hay et al. 2008). Some have used this result to criticize tuataras' status as a living fossil (e.g., Carnall 2016). Turner (2019)'s first criticism of this inference is that the Hay et al. (2008) study only uses mitochondrial DNA, which

would not be expected to influence morphology (14). Turner's next point is more relevant for our purposes: he says that even if the study *had* used nuclear DNA, "developmental processes might insulate morphology from rapid molecular change" and that "[r]apid molecular change in the nuclear genome could also reflect selection pressures on aspects of the organism, like the immune system, that never show up in the fossil record" (14). These criticisms of the skeptics of tuataras' living fossil status line up nicely with my evaluation of the inference from morphological stability to genetic stability to a slow evolutionary rate. Turner concludes that "in spite of the high rate of molecular change, tuataras are a clear instance of a phylogenetic living fossil taxon" (15).[4]

However, Turner (2019) does *not* say that we can save the tuataras' living fossil status by appealing to a different idea of evolutionary rates. Indeed, his reconstruction of the argument against tuataras counting as living fossils is that "living fossils must have especially slow rates of evolutionary change, whereas the molecular evidence points toward especially rapid evolution in tuataras" (14). Turner's criticisms of this line of reasoning challenge the idea that a slow rate of evolution is a necessary feature of a living fossil taxa, rather than the idea that a slow rate of molecular change may not line up with a slow rate of (character) evolution at all. Later, in discussing coelacanths (another candidate for a living fossil taxon), Turner references "rates of morphological change," but does not equate these rates with rates of evolutionary change (16). Indeed, Turner says that "morphological stability in certain characters is entirely compatible with evolutionary change happening under the geological radar" (18). However, as I have

[4]Interestingly, Hay et al. (2008) also interpret their results about the faster-than-expected rate of molecular evolution in tuataras as being evidence that "rates of neutral molecular and phenotypic evolution are decoupled" (106).

argued, morphological stability in certain characters is exactly *not* compatible with evolutionary change happening *on those characters*. My suspicion is that Turner is confusing rates of evolutionary change with rates of molecular change here.

Werth and Shear (2014) have a similar take on the case of tuataras. While Werth and Shear do not think that evidence of a higher rate of molecular evolution in this lineage disqualifies it as a living fossil taxon, they say that the high molecular rates "provide strong evidence countering the misconception that living fossils have stopped evolving" (438). In other words, Werth and Shear – like Turner – apparently want to maintain the tuataras' status as a living fossil by arguing that living fossils need not have a slow rate of evolutionary change, rather than by claiming that rates of evolutionary change are best measured at the level of traits and not genes, necessarily (although insofar as genes are themselves traits, a rate of evolution could apply to them as well).[5]

One implication of focusing on the inference from morphological stability to slow rates of evolutionary change is that it is not clear what the epistemic role evidence of *molecular* stability in a lineage could have. Lidgard and Love (2018) say, "the primary role of the living fossil concept is to mark out more precisely what requires explanation in a given instance for a particular entity in order to account for morphological *and molecular* stability or persistence over long periods of evolutionary time" (763, emphasis added). If molecular stability does not let us infer an evolutionary rate (other than an evolutionary rate at the molecular level itself), then why might we want to know about

[5]Note that Werth and Shear do acknowledge that "Some biologists speculate that mere genetic change does not translate to evolutionary change" and that there is "independence between molecular and morphological evolution," although they do not endorse this position (439).

molecular stability at all? Lidgard and Love have (at least) one interesting response: we might want to know how molecular and morphological rates of change are related or decoupled (766).

It is especially interesting that the inference to slow evolutionary rates from morphological similarity is the most secure of those I have considered in this paper, because Darwin (1859/1964)'s use of the term "living fossils" was in the context of explaining why some lineages display slower rates of evolutionary change than others. While Darwin's explanation was that these lineages had been "exposed to less severe competition" (107), and now we know that the reasons for stabilizing selection are more complicated, he still made, by my account, the most reasonable inference from the morphological similarity of extant and past taxa.

The various attempted inferences and critiques of these inferences examined in this section are summarized in table 1.

4 Conclusion

This paper's primary contribution has been to disambiguate the inferences that we can justifiably make on the basis of classifying a taxon as a living fossil. In doing so, I have specified some of the ways in which the living fossil concept may be epistemically useful. This adds to claims that the living fossil concept is epistemically useful in other ways, such as by identifying phenomena in need of explanation. I also intend to complement, not supplant, accounts in which the living fossil concept is useful for non-epistemic

F	Inference	Evaluation
General phenotypic similarity	morphological similarity → genetic similarity → general phenotypic similarity	Morphological similarity does not imply genetic similarity
General phenotypic similarity	morphological similarity → general phenotypic similarity	Morphological similarity only implies phenotypic similarity for some phenotypes (developmental modularity)
Persistence of lineage	morphological similarity → genetic similarity → persistence of lineage	Morphological similarity does not imply genetic similarity
Persistence of lineage	morphological similarity → persistence of lineage	Morphological similarity does not imply persistence of lineage, but it might be the best evidence we have
Slow evolutionary rate	morphological similarity → morphological stability → genetic stability → slow evolutionary rate	Morphological stability does not imply genetic stability, and genetic stability does not imply a slow evolutionary rate
Slow evolutionary rate	morphological similarity → morphological stability → slow evolutionary rate	Morphological stability *does* imply a slow evolutionary rate, relative to that morphology

Table 1: Summary.

reasons. One possible area for future research is identifying the ways in which the epistemic and non-epistemic uses of the concept may interact. For instance, Turner thinks that a living fossil taxon's high contribution to phylogenetic diversity has implications for conservation efforts. But we may need to address epistemic issues before we are able to draw appropriate normative conclusions.

This paper has also served as an example of how developmental biology can be useful for paleontologists. Historically, development hasn't been given much consideration in making claims about fossils, largely because fossil evidence does not include information about developmental processes. Discussions of homology in general, which are relevant

to persistence of lineage, involve the contributions of both paleontologists and developmental biologists. Living fossils serve as another good example for how considerations from developmental biology and paleontology could be productively combined, because we have evidence about fossilized as well as living taxa. The arguments I have made in this paper, such as those regarding developmental modularity, may have other implications for paleontology outside of the context of living fossils, and more generally point to the fertility of of exploring the intersection between developmental biology and paleontology.

References

Carnall, M. 2016. "Let's make living fossils extinct". *The Guardian*. https://www.theguardian.com/science/2016/jul/06/why-its-time-to-make-living-fossils-extinct.

Casane, D., and P. Laurenti. 2013. "Why coelacanths are not "living fossils"". *Bioessays* 35:332–338.

Currie, A. M. 2016. "The mystery of the triceratops mother: How to be a realist about the species category". *Erkenn* 81:785–816.

Darwin, C. 1859/1964. *On the origin of species*. Cambridge, MA: Harvard University Press.

Fortey, R. 2011. *Horseshoe crabs and velvet worms: The story of the animals and plants that time has left behind*. New York, NY: Vintage Books.

Fusco, G., and A. Minelli. 2010. "Phenotypic plasticity in development and evolution: facts and concepts". *Philosophical Transactions of the Royal Society B* 365:547–556.

Gilbert, S. F. 2000. *Developmental Biology*. Sunderland, MA: Sinauer Associates.

Hay, J. M., et al. 2008. "Rapid molecular evolution in a living fossil". *Trends Genet* 24 (3): 106–109.

Ho, S. 2008. "The molecular clock and estimating species divergence". *Nature education* 1 (1): 168.

Jablonski, D., and N. H. Shubin. 2015. "The future of the fossil record: Paleontology in the 21st century". *PNAS* 112 (16): 4852–4858.

Laland, K. N., et al. 2015. "The extended evolutionary synthesis: its structure, assumptions and predictions". *Proc. R. Soc. B.* 282:20151019.

Lidgard, S., and A. C. Love. 2018. "Rethinking living fossils". *Bioscience* 68 (10): 760–770.

Mathers, T.C., et al. 2013. "Multiple global radiations in tadpole shrimps challenge the concept of 'living fossils'". *Peer J* 1:e62.

Nijhout, H. F. 2019. "The multistep morphing of beetle horns: Genes that specify insect wings initiate horn development in dung beetles". *Science* 366 (6468): 946–947.

Omland, K. E., L. G. Cook, and M. D. Crisp. 2008. "Tree thinking for all biology: The problem with reading phylogenies as ladders of progress". *BioEssays* 30:854–867.

Schopff, T. J. M. 1984. "Rates of evolution and the notion of "living fossils"". *Ann. Rev. Earth Planet. Sci.* 12:245–292.

Turner, D. D. 2016. "A second look at the color of dinosaurs". *Studies in History and Philosophy of Science* 55:60–68.

— . 2019. "In Defense of Living Fossils". *Biology & Philosophy* 34 (23).

— . 2011. *Paleontology: a philosophical introduction.* New York, NY: Cambridge University Press.

Wagner, P., et al. 2017. "Ontogenetic sequence comparison of extant and fossil tadpole shrimps: no support for the "living fossil" concept". *PalZ* 91:463–472.

Werth, A. J., and W. A. Shear. 2014. "The evolutionary truth about living fossils". *American Scientist* 102:434–443.

West-Eberhard, M. J. 2003. *Developmental Plasticity and Evolution*. New York, NY: Oxford University Press.

— . 2019. "Modularity as a universal emergent property of biological traits". *Journal of Experimental Zoolology* 332:356–364.

— . 2005. "Phenotypic Accommodation: Adaptive Innovation Due to Developmental Plasticity". *Journal of experimental zoology* 304B:610–618.

This is an early draft to be presented at: The 27th Biennial Meeting of the Philosophy of Science Association, 2020, Baltimore, MD. Please do not cite or quote without permission. Any criticisms or suggestions are welcome. Contact Email: tungyingwu@outlook.com

Structural Decision Theory

Tung-Ying Wu
Fudan University
tungyingwu@outlook.com

Draft as of March 5, 2020

Abstract

Judging an act's causal efficacy plays a crucial role in causal decision theory. A recent development appeals to the causal modeling framework with an emphasis on the analysis of intervention based on the causal Bayes net for clarifying what causally depends on our acts. However, few writers have focused on exploring the usefulness of extending structural causal models to decision problems that are not ideal for intervention analysis. I found that it is structural models, rather than intervention analysis, serves as a valuable formal tool for a range of realistic decision problems that involve mixed causal mechanisms. The thesis concludes that structural models provide a more general framework for rational decision-makers.

1. Introduction

Decision theories concern an agent's rational choice in a decision problem, where the agent faces different acts to choose from but is uncertain about each act's possible consequences. Suppose she knows about the possible consequences of her different acts, the utility of each consequence, and the probability of each consequence. Then she can

Draft-Do Not Cite

acquire the expected utility of each act by multiplying the probability and the utility of each possible consequence of an act, and then adding the results of all possible consequences of the act. Philosophers in decision theory contend that a rational choice for an agent is an option that maximizes expected utility.

Causal decision theory (hereafter, CDT) endorses the principle of expected utility maximization, but holds that the agent must take the causal relevance of her acts to their outcomes into consideration. Proponents of CDT share the belief that rational agents should maximize expected utility based on the causal information relevant to their acts, but differ in what approach best captures an act's causal efficacy.[1]

Interventionist decision theory (hereafter, IDT) is a form of CDT because IDT also holds that the relevant information that matters to our decision should be causal, but IDT approaches an act's causal efficacy through intervention analysis within the framework of causal modeling.[2] Specifically, IDT holds that an agent should conceive of an act as an

[1] David Lewis, 1981, p. 11; James Joyce, 1999, pp. 146; Ralph Wedgwood, 2013, p. 2644; Arif Ahmed, 2014, pp. 8-9; Paul Weirich, 2016.

[2] Peter Spirtes, Clark Glymour, and Richard Scheines (2000, pp. 47-53) and Judea Pearl (2009, pp. 23-4, 70-4) claim that an intervention I as an external force sets X to certain values, and I neither causes any variable other than X nor is caused by any other variable in a causal model.

More formally, intervention analysis is assessed by the theory of causal Bayes net. Variables (denoted by uppercase letters) represent tokens of events that serve as relata of (type level) causal relations, and these variables range over possible values (denoted by lowercase letters) that represent these events' occurrence or non-occurrence, or a value if an event is of a quantity. A Bayesian causal model M is a triple $<G, V, P>$, where V is a set that contains variables whose causal relationships we are interested in studying, P is the probability distribution of each variable, and G is a directed acyclic graph. G consists of nodes that represent variables in M, and arrows between nodes that represent causal relations. If the value of a variable Y depends on X, then there will be a directed path from X to Y. P satisfies the causal Markov condition if and only if each variable X_i in V is independent of all other variables except X_i's descendent given X_i's parent PA_i, where "X_i's descendent" stands for the other variables in V that are causally downstream from X_i and "X_i's parents" stand for X_i's immediate causes. More specifically, P satisfies the causal Markov condition if and only if the following condition holds: $P(X_1, \ldots, X_n) = \prod_i P(X_i|PA(X_i))$, where $X_1, \ldots, X_n$ are all variables in V, and "PA_i" stands for "parents of X_i."

An intervention on X_j removes all its pre-existing cause and set it to a specific value. Hence, the intervention analysis is done by removing $P(X_j|PA(X_j))$ from the above joint distribution. This amounts to set X_j to a specific value and make it no longer depends on its original parents. Hence, the effect of

Draft Do Not Cite

intervention that disables all pre-existing causes of the act in a decision problem.[3,4] This is because causal models represent the causal details relevant to a decision-making context in a rigorous mathematical language. Hence, when engaging with a decision problem, one should use causal models to clarify one's assumptions about the causal structure of the problem, the information that one has available, and the question one is asking. More importantly, by making use of causal models, one can distinguish causation from correlation.[5]

IDT instructs rational agents to choose an act x that maximizes the interventionist expected utility (hereafter, IEU. See below.). Let Y be a random variable that ranges over possible outcomes, P be a rational agent s's subjective probability function, $do(X = x)$ be s's intervention to make s do x, $V(Y = y)$ be the utility of an outcome y, and $IEU(x)$ be the interventionist expected utility of act x.[6] Here is Pearl's definition of IEU:[7]

$$IEU(x) =^{\text{df}} \sum_y P(Y = y \mid do(X = x))\, V(Y = y)$$

This definition asserts that s should assess the expected utility of an outcome y based on evaluating the effect of the intervention to make s do x.

---

intervention on X_j is obtained by the new joint distribution: $P'(X_1, \ldots, X_n) = \prod_{i \notin j} P(X_i | PA(X_i))$.

[3] See Christopher Meek and Clark Glymour, 1994, pp. 1007-8; Pearl, 2009, p. 70 and pp. 108-112; Christopher Hitchcock, 2016, pp. 1158-9; Reuben Stern, 2017, pp. 4139-42; Stern, 2018, pp. 2-3. Meek and Glymour (1994) claim that we may conceive of our acts as interventions only when we believe that our actions are not caused by circumstances beyond our control. See Hitchcock, 2016, p. 1166, and Stern, 2018, pp. 7-8.

[4] Note that the notion of "intervention" in this paper is not the same as James Woodward's (2003, pp. 94-98). In this paper, "intervention analysis" is understood in terms of manipulating the probability distribution in a causal model where the causal Markov condition holds. See footnote 2.

[5] Meek and Glymour, 1994; Pearl, 2009, section 4.1; Hitchcock, 2015, p. 1175; Stern, 2017, p. 4147.

[6] Pearl uses the do-operator to denote "intervention."

[7] Pearl, 2009, p. 108. For similar proposals, see Meek and Glymour, 1994, pp. 1009-10; Hitchcock, 2016, pp. 1162-4.

Draft Do Not Cite

Nevertheless, Pearl (2017 and forthcoming) recently proposes a new definition of expected utility in terms of structural causal models (hereafter, SCM) as decision-making conditionals. Call the definition of expected utility with an application of SCM "the structural expected utility" (hereafter, SEU):[8]

$$SEU(x) =^{\text{df}} \sum_y P(Y_x = y)\ V(Y = y)$$

Pearl entitles $P(Y_x = y)$ as a SCM defined counterfactuals.[9] This definition declares that *s* should evaluate the expected utility of act *x* by using an SCM analysis of causality.

IEU and SEU are methodologically different approaches. They instruct the agent to use different procedures in evaluating the causal information of decision problems. For instance, IEU tells the agent to obtain the probability distribution and the corresponding causal graph of each variable in a decision problem.[10] In contrast, SEU requires delineating functional relations between relevant variables to attain the causal structure.[11] They are nevertheless different methodologies for the agent to approach decision problems.

This paper attempts to assess the scope of SEU and IEU, their effectiveness in making explicit the causal structure of decision problems. Previous work has only focused on IEU's implications for some controversial examples in CDT, such as

[8] Pearl, 2017, p. 1 and forthcoming, p. 1. Note that Pearl (2009, p. 108) originally endorsed IEU. Also, Pearl sometimes uses $P(Y = y \mid do(X = x))$ and $P(Y_x = y)$ interchangebly in his writings because the later can be translated and computed by the former under several strong assumptions. Such translation would fail in some examples. See Pearl, 2009, pp.245-7, 289-93 and Pearl et al., 2016, pp. 107-116.

[9] Pearl, forthcoming, pp. 2-6. For the comparison between the causal modeling's and Lewis's accounts of counterfactuals, see Eric Hiddleston, 2005, Woodward, 2003, pp. 133-145, and Pearl, 2009, pp. 238-41, and Pearl, 2017.

[10] See footnote 2.

[11] I will formally expand on this later.

Draft Do Not Cite

Newcomb's Problem and Psychopath Button, or issues of uncertainty about causal dependency.[12] To the best of my knowledge, the distinction between IEU and SEU has not been dealt with in depth. The example in next section demonstrates that it is SEU, rather than IEU, serves as a valuable formal tool for a range of realistic decision problems that involve mixed causal mechanisms. Therefore, SEU provides a more general framework for rational decision-makers.

The following sections of this paper are organized as follows. Section 2 presents the example of the Spinner and explains why IEU fails to deliver an intuitive result. Section 3 gives a brief overview of SCM. Section 4 employs SCM to analyze the Spinner and shows how SCM and SEU, but not intervention analysis and IEU, deliver an intuitive result.

2. The Spinner

An agent has a chance to win a prize (called the reward). There is a spinner (drawn below) and an arrow in the circle. The agent may choose between two options "SAFE" and "ADD-X." If the agent plays SAFE, the agent flicks the arrow and gains the value where the arrow stops. Since 40% of the time the arrow stops in area Z =1, 20% in area Z = 2, and 40% in area Z = 3, the expected average gain for the agent is 2 units of money. In contrast, option ADD-X allows the agent to increase the reward by X unit(s) of money for a small cost (much smaller than X) with the following rule: if the arrow stops in area Z = 1, Z will not be contributive, and the reward will have only X unit(s) of money. If the arrow stops in area Z = 3, Z will be contributive so the reward will have 3 +

[12] Meek and Glymour, 1994, pp. 1008-9; Hitchcock, 2016, pp. 1165-9; Stern, 2017, pp. 4142; Stern, 2018, pp. 15-16.

Draft Do Not Cite

X units. However, if the arrow stops in area $Z = 2$, Z will be deleterious so the reward will have $X - 2$ units.

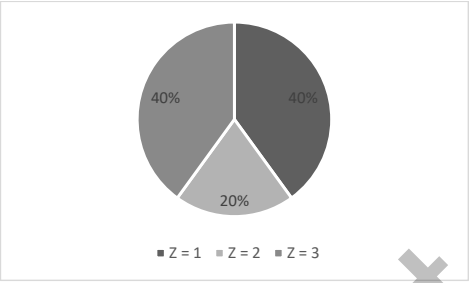


Figure 1. The Spinner

Now, assessing the expected gain of option ADD-X is a complicated task.[13] The spinner is a mixture of areas of Z that react differently to the agent's choosing ADD-X. For example, Z is contributive to the reward in $Z = 3$, not contributive to the reward in area $Z = 1$, and deleterious to the reward in area $Z = 2$. The causal mechanisms of these areas differ from area to area because they exhibit different dispositions that manifest given the presence of the agent's acting on ADD-X.

Since the spinner consists of the areas with different dispositional properties, the intervention analysis has difficulty in accurately predicting the causal effect of choosing ADD-X. Simply put, intervention analysis is mostly assessed by the theory of causal Bayes net with the assumption that the relevant causal model satisfies the causal Markov condition. However, the procedure of computing the effect of acting on ADD-X as an

[13] This example is a modified case of "additive intervention." Namely, one evaluates the effect of adding some amount from X without removing a pre-existing causal process of X. (In Newcomb's Spinner, I use X as an instrument variable, Y as some amount, Z as a preexisting cause of Y.) See Bill Shipley, 2016, pp. 9-11, 50-4 and Pearl, 2009, section 11.4.4. Pearl et al. (2016, pp. 109-111) confirm that the effect of additive intervention could not be reduced to intervention-expressions alone.

Draft Do Not Cite

intervention amounts to computing $P(Y = y \mid do(X = q))$, which does not fix the level of Z. Since the level of Z is not fixed, we may estimate the value of Z by the expectation ($E(Z)$), and $E(Z) = 2$ in the Spinner. Thus, the intervention analysis implies that the agent should predict that acting on ADD-X as an intervention will always result in the worst case scenario: Z will be deleterious so the reward will have $X - 2$ units. Nevertheless, this is certainly incorrect. For only 20% of the time the value of Z is deleterious to the reward, but 80% of the time the value of Z is not deleterious to the reward. It seems that ADD-X does not always lead to the worst causal scenario of the value of Z being deleterious. Intervention analysis is limited when it is not possible for the agent to intervene on a relevant feature that has a mixture of different causal mechanisms.[14]

In the Spinner, the agent cannot intervene to fix the amount of the reward. For doing so is an intervention that removes the pre-existing rule of the spinner, but the agent must flick the arrow, and it is not up to the agent to fix the arrow on the spinner that consists of areas that react differently to adding X. Thus, it seems that the intervention analysis of choosing ADD-X is unfitting if the intervention analysis is insensitive to the variant causal properties across the circle that is not intervenable. Hence, the agent's intervention analysis of choosing ADD-X is inaccurate, and it remains unclear whether the agent should choose SAFE or ADD-X.

How do we evaluate the causal efficacy of an act when the world is a mixture of variant mechanisms in which the act causes different outcomes? Presumably, if the agent

[14] One cannot evaluate the causal efficacy of ADD-X by the analysis of interventions $P(Y = y \mid do(X = x, Z = z))$, $P(Y = y \mid do(X + Z))$, and $P(Y = y \mid do(X - Z))$. As stipulated in the example, it is not possible for the agent to intervene to set Z to a fixed value.

Draft Do Not Cite

knows each area's causal mechanism, she should evaluate the causal effects of her interventions area by area. Since the issue is predicting the expected gain of ADD-X, the agent should average the causal effects in each area by its proportion to the whole circle to derive the desired quantity.

This paper puts forward a justification for applying SCM to evaluate an act's causal efficacy in decision theory. The last question of the above example—how we evaluate the causal efficacy of an act when the world is a mixture of variant mechanisms in which the act causes different outcomes—calls for a SCM analysis. For this purpose, the above example provides an independent reason for employing SCM to define an act's expected utility in decision theory, namely, SEU. In what follows, I will introduce SCM, which may be of use to a rational agent to accurately predict what is causally downstream from her acts.

3. Structural Causal Models

SCM can formally represent causal relations in a rigorous mathematical language. They conveniently represent an agent's belief about causal relationships among variables of interest and the causal effect of an intervention. A prior development of SCM includes the work of the economist Herbert A. Simon who specialized in decision-making. In his influential papers, Simon argues that we can define a causal system as some functional relationships in a structure—a specific arrangement of variables and equations in fixing the sequence of computing their solutions.[15] I will begin with a brief account of SCM.

[15] Herbert A. Simon, 1957, pp. 10-13.

Draft Do Not Cite

A structural causal model M consists of a quadruple $<U, V, f, P>$, where U is a set of exogenous (or background) variables, V is a set of endogenous variables. Exogenous variables represent background factors in M and are only determined by factors outside the model, and their values do not depend on the other variables in the model. In contrast, endogenous variables are determined only by the other variables in the model. f is a set of functions that assign each endogenous variable in V a value based on the values of the other variables in the model. P represents a probability distribution over all variables in U. Specifically, each function has the form:[16]

$X_i = f_i (PA_i, U_i), i = 1,\ldots,n$

where X_i is an endogenous variable in V, PA_i (which stands for "parents of X_i") is a set of variables in V, U_i is an exogenous variable in U, and PA_i and U_i together determine the value of X_i. Moreover, by assumption, each variable in V can only have one distinct equation that determines its value. Hence, each function represents an autonomous causal mechanism that predicts what value nature would assign to X_i in response to every possible value combination of (PA_i, U_i). They are autonomous in the sense that one function f_i continues to hold or remains undisrupted by external changes to the other functions in f. Hence, the causal relations in M are deterministic given a value assignment of U_i. Since every X_i is (partially or wholly) determined by at least one U_i and every U_i is not determined by any X_i in V, a value assignment of all U_i in U determines a unique value distribution over all X_i in V based on f. If P is the probability distribution

[16] Simon, 1957, pp. 18-19, 40; Pearl, 2009, pp. 202-3; Pearl et al, 2016, pp. 26-7.

Draft Do Not Cite

over all exogenous variables, the probability distribution for the endogenous variables is also P.[17, 18]

A structural causal model M corresponds to a causal graph G. If the value of a variable Y depends on X according to the function f_Y, then there will be a directed path from X to Y.[19]

For the sake of illustration of SCM and an explicit representation of the causal relationships in the Spinner, I will use SCM to represent the Spinner, and demonstrate that the agent can accurately predict what is causally downstream of her acts in SCM's expressions in the next section.

4. Additive Intervention

I use the linear SCM M_1: $\langle U, V, f, P \rangle$ to represent the causal relationships in the Spinner. Let X, Y, Z be endogenous variables in V, and I, U_Z be exogenous variables in U.[20] These variables range over possible values (denoted by lowercase letters). X represents how much value the agent adds to the prize, Y represents the value of the reward, and Z represents the value that the arrow points to. The intervention variable I represents the agent's intervention, and it is an exogenous variable because only outside factors (for example, the agent's free will) determine its value.

[17] Simon, 1957, pp. 40-3, 54-6; Pearl, 2009, pp. 27-32, 205-6; Pearl et al, 2016, p. 98.
[18] A consequence of a structural causal model M is that the probability distribution of every variable in M satisfies the causal Markov condition (CMC). CMC holds in SCM under these further assumptions: (a) there is no causal loop in M, namely, the associated causal graph is acyclic; (b) the exogenous variables in U are jointly independent; (c) M includes every variable that is a cause of two or more other variables; and (d) if any two variables are dependent, then one is a cause of the other or there is a third variable causing both. See Pearl, 2009, p. 30; Daniel Steel, 2005, p. 10.
[19] Pearl, 2009, p. 203.
[20] For the sake of brevity, I omit some exogeneous variables.

Draft Do Not Cite

In the Spinner, X can increase the prize Y, and Z also causally affects Y's value. The causal graph G_1 of this model M_1 is figure 2:

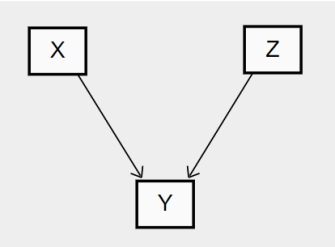


Figure 2. The causal graph G_1 of the Spinner with some exogenous variables omitted.

The following functions represent the causal relations between these variables:

f_X: $X = \{q, 0\}$

f_Z: $Z = U_Z$

$$f_Y\text{: } Y = \begin{cases} Z \; if \; X = 0 \\ X \;\; if \; X > 0 \; and \; Z < 2 \\ X - Z \;\; if \; X > 0 \; and \; Z = 2 \\ Z + X \;\; if \; X > 0 \; and \; Z \; > 2 \end{cases}$$

U_Z is an exogenous variable that determines the value of Z. The probability distribution of U_Z is the composition of the spinner: $P(U_Z = 1) = 0.4$, $P(U_Z = 2) = 0.2$, and $P(U_Z = 3) = 0.4$. Also, $X = q$ represents the agent's action to add q to the reward; $X = 0$ represents the agent's action to add nothing to the reward.

Next, f_X stands for the causal mechanism that specifies how the agent decides to add value to the reward: if she decides to add q amount of value, X will be set to q. If she decides to add nothing, X will be set to 0.

f_Y stands for how X and Z determine the amount of the reward: if the agent adds no value ($X = 0$), then the value of Y will equal Z. If the agent adds some value ($X > 0$), but Z is lower than 2, then the value of Y will be X. If the agent adds some value, but Z equals

Draft Do Not Cite

2, then the amount of Y will be $X - Z$. If the agent adds some value, and Z is larger than 2, then the value of Y will be $Z + X$. This function f_Y demonstrates different mechanisms in which Z reacts differently to the added value X in the process of determining the reward Y.

Turning now to the question of how an agent predicts the overall causal effects of choosing ADD-X. The diverse areas have varied types of causal mechanisms represented by several levels of Z. In the Spinner, the circle consists of areas with three levels of Z: 40% is Z =1, 20% is Z =2, and 40% is Z = 3. One can estimate the results in each level of Z and averages these effects by the probability distribution of Z.[21] I now turn to explain how this sort of prediction is done in M_1.

One may use $P(Y_x = y | Z = z)$ to represent the probability that an outcome y would obtain conditional on the action $X = x$ in a structural model updated by $Z = z$.[22] Given a structural model M and observed information $Z = z$, one can evaluate the conditional $P(Y_x = y \mid Z = z)$ in three steps:[23, 24]

(1) Abduction: Conditionalize on the evidence z to determine the value of the variables in U.
(2) Action: Replace the equations corresponding to variables in set X by the equation $X = x$.
(3) Prediction: Use the modified model and the updated value of the variables in U to compute the value of Y.

[21] In cases where experimental units manifest variant dispositional properties, Spirtes, *et al*. (2000, p.165-7) also cite similar calculations to obtain predictions.
[22] $P(Y_x = y)$ is a subjunctive conditional. "$P(Y_x = y)$" stands for the probability that, had an intervention $do(X = x)$ been performed, an outcome $Y = y$ would obtain.
[23] The following procedure draws from David Galles and Pearl (1998), Pearl (2009, pp. 202-6), and Pearl et al. (2016, pp. 92-8). Joseph Halpern (2000) provide another detailed account of causal inferences in SCM. For the sake of simplicity, I will skip some unnecessary technical details. Note that this is different from Woodward's notion of causality analyzed with counterfactual interventions.
[24] Pearl, 2009, p. 37, 206.

Draft Do Not Cite

The first step uses the information $Z = z$ about the situation to fix the values of the exogenous variables in U. In particular, each value assignment of variables in U is the defining characteristic of a single individual or situation. For example, in the model M_1, a value assignment $U_i = u_i$ stands for the identity of the agent and the spinner. The second step stands for the minimal modification of the model M that replaces f_X with $X = x$. The third step predicts the value of Y based on the modified M and the updated values of U.

Returning to the question posed in the Spinner, it is now possible to answer the agent's question of assessing SEU of choosing ADD-X by SCM. First, the agent updates her value assignment of U from the supposition that $Z = 1$, 2, or 3 and identifies U_Z. Next, she carries over the updated value of U_Z to the model M_1 modified by $X = q$. Finally, she predicts the value of Y by finding a solution to the following equations:

f_X: $X = q$

f_Z: $Z = U_Z$

$$f_Y\text{: } Y = \begin{cases} Z \; if \; X = 0 \\ X \; if \; X > 0 \; and \; Z < 2 \\ X - Z \; if \; X > 0 \; and \; Z = 2 \\ Z + X \; if \; X > 0 \; and \; Z > 2 \end{cases}$$

Next, she can predict that had she added q unit(s) to the reward when $Z = 1$, the reward would be q. Equally, she can also predict that had she added q unit(s) to the reward when $Z = 2$, the reward would be $q - 2$. Had she added q unit(s) to the reward when $Z = 3$, the reward would be $q + 3$. Given that 40% of the time $Z = 1$, 20% of the time $Z = 2$, and 40% of the time $Z = 3$, the SEU of "ADD X" would be $q + 0.8$ minus the fee that the agent has to pay. Recall that the expected value of the reward if the agent plays Safe is invariably 2. Therefore, if option ADD-X allows the agent to pay less than 0.1

Draft Do Not Cite

unit of money to add $X > 1.3$ to the reward, she will be quite confident that option ADD-X is preferable to option SAFE.

The implication is that facilitating SCM and deriving SEU in the Spinner and similar situations is more fitting than intervention analysis. As demonstrated in the Spinner, the approach of SCM captures the mixture of variant causal mechanisms specified by the probability distribution of Z and the function f_y, and thereby obtains more accurate characterizations of each area's causal property and the causal efficacy of choosing ADD-X. Hence, in cases where an agent observes different causal properties that are not intervenable across the population in the real world, the agent might more adequately make statements about her acts' causal efficacy in SCM's mathematical terms.

The cases of mixed causal properties are realistic, but often not ideal for intervention analysis that is appropriate when most members of a population share invariant causal profiles. These cases are common when an act causally affects an extensive system. For example, a socioeconomic policy affects diverse citizens; an educational program affects numerous students; a business decision affects countless customers; an approved drug affects various patients. It would seem that these complicated situations are not rare in decision problems.

In this paper, I have identified the example of the Spinner which underlines the importance of SCM and SEU. In that example, the characterization of the causal effect of the act delivered by IEU and the characterization delivered by SEU diverge and the latter—not the former—seems intuitively correct. Moreover, the language of SCM and SEU is richer than intervention analysis and IEU because SCM and SEU enable the agent

Draft Do Not Cite

to make necessary mathematical statements that relate directly to various causal dispositions in the real world.[25] The theoretical implication of the Spinner is that SEU is recommended in similar situations, and that SEU might be a foundation for a more general decision theory.

References

Ahmed, Arif. 2014. *Evidence, Decision and Causality*. Cambridge University Press.
Dawid, Philip. 2015. "Statistical Causality from a Decision-Theoretic Perspective." *Annual Review of Statistics and Its Application* 2: 273–303.
Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116 (1): 93–114.
Galles, David, and Judea Pearl. 1998. "An Axiomatic Characterization of Causal Counterfactuals." *Foundations of Science* 3 (1): 151–182.
Gibbard, Allan, and William Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, edited by A. Hooker, J. J. Leach, and E. F. McClennen, 125–62. D. Reidel.
Halpern, Joseph Y. 2000. "Axiomatizing Causal Reasoning." *Journal of Artificial Intelligence Research* 12 (1): 317–337.
Hiddleston, Eric. 2005. "A Causal Theory of Counterfactuals." *Noûs* 39 (4): 632–57.
Hitchcock, Christopher. 2013. "What Is the 'Cause' in Causal Decision Theory?" *Erkenntnis (1975-)* 78: 129–46.
———. 2016. "Conditioning, Intervening, and Decision." *Synthese* 193 (4).
Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press.
Lewis, David. 1981. "Causal Decision Theory." *Australasian Journal of Philosophy* 59 (1): 5–30.
Meek, Christopher, and Clark Glymour. 1994. "Conditioning and Intervening." *The British Journal for the Philosophy of Science* 45 (4): 1001–21.
Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2nd edition. Cambridge, U.K. ; New York: Cambridge University Press.
———. 2017. "Physical and Metaphysical Counterfactuals: Evaluating Disjunctive Actions." *Journal of Causal Inference* 5 (2): 1–10.
Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. 1 edition. Chichester, West Sussex: Wiley.
Shipley, Bill. 2016. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R*. 2 edition. Cambridge: Cambridge University Press.

[25] Philip Dawid (2015, pp. 280-2) considers several formal frameworks for analyzing causal processes in decision problems. He agrees that intervention-expressions are not as flexible as the language of SCM.

Draft Do Not Cite

Shpitser, Ilya, and Judea Pearl. 2009. "Effects of Treatment on the Treated: Identification and Generalization." In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 514–521. UAI '09. Arlington, Virginia, United States: AUAI Press

Simon, Herbert Alexander. 1957. *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Society Setting*. New York: John Wiley and Sons.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search, Second Edition*. Second edition. Cambridge, Mass: A Bradford Book.

Steel, Daniel. 2005. "Indeterminism and the Causal Markov Condition." *British Journal for the Philosophy of Science* 56 (1): 3–26.

Stern, Reuben. forthcoming. "Decision and Intervention." *Erkenntnis*, 1–22.

———. 2017. "Interventionist Decision Theory." *Synthese* 194 (10): 4133–4153.

Wedgwood, Ralph. 2013. "Gandalf's Solution to the Newcomb Problem." *Synthese* 190 (14): 2643–75.

Weirich, Paul. 2016. "Causal Decision Theory." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University.

Woodward, James. 2005. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, USA.

Draft Do Not Cite

The Exploratory Role of Explainable Artificial Intelligence

August 2020

Carlos Zednik
carlos.zednik@ovgu.de

Hannes Boelsen
hannes.boelsen@ovgu.de

Otto-von-Guericke-Universität Magdeburg

To be presented at the 27th Biennial Meeting
of the Philosophy of Science Association

Abstract

Models developed using machine learning (ML) are increasingly prevalent in scientific research. Because many of these models are opaque, techniques from Explainable AI (XAI) have been developed to render them transparent. But XAI is more than just the solution to the problems that opacity poses—it also plays an invaluable exploratory role. In this paper, we demonstrate that current XAI techniques can be used to (1) better understand what an ML model is a model *of*, (2) engage in causal inference over high-dimensional nonlinear systems, and (3) generate algorithmic-level hypotheses in cognitive science.

Acknowledgements

This work was supported by the German Research Foundation project ZE-1062/4-1. Carlos Zednik is thankful for the opportunity to attain insights into AI industry applications at the neurocat GmbH in early 2020.

1. Introduction

Models developed using machine learning ("ML models") are increasingly prevalent in scientific research. In neuroscience, ML-programmed classifiers are used to specify the representational contents of brain states and to predict human behavior from fMRI data (Ritchey et al. 2017). In astrophysics, classifiers trained on telescope imagery are used to determine the possible location of exoplanets (Datillo et al. 2019). In materials science, machine learning is used to discover stable materials and to predict their crystal structure (Schmidt et al. 2019).

Recent discussions have focused on the fact that many ML models are *opaque* (Humphreys 2009). Loosely speaking, a model is opaque when it is difficult to understand why it does what it does or to know how it works. Recent attempts to assess the impact of opacity generally agree that opacity prevents different stakeholders[1] from achieving goals such as intervening on the system when it breaks down, or evaluating its behavior against ethical and legal norms (Burrell 2016; Hohman et al. 2018; Zednik 2019).

In philosophy of science, the most important stakeholder is the *scientific investigator*. Scientific investigators are known to use ML models to achieve epistemic goals such as describing a phenomenon (e.g., distinguishing the fMRI signatures of fear and excitement), predicting new observations (e.g., determining the probable location of an exoplanet), and explaining observed data (e.g., identifying a causal link between smoking and lung cancer). Opacity can negatively impact scientific research by preventing investigators from using ML models to achieve some or all of these epistemic goals.

That said, little is known about the positive impact of recent attempts to overcome opacity through *Explainable Artificial Intelligence* (XAI). This nascent research program aims to develop analytic techniques with which to render opaque models *transparent* by answering questions about why they do what they do or how they work.[2] Whereas these techniques' importance for industry and governance is becoming increasingly apparent (Doran et al. 2017; Wachter et al. 2018), their utility for scientific research remains uncertain.

This paper argues that Explainable AI can play an invaluable but hitherto unrecognized role in *scientific exploration*. Recent discussions of exploration distinguish at least four distinct but not mutually exclusive aspects (for discussion see e.g., Gelfert 2016): identifying a *starting point* for future inquiry; providing a *proof-of-principle* demonstration; providing a *potential explanation* of a specific (type of) phenomenon; and assessing the *suitability of a particular target*. Whereas previous contributions have considered the exploratory role of ML models in their own right (e.g., Cichy & Kaiser 2019), little is known about the unique exploratory utility of Explainable AI.

1 Tomsett et al. (2018) provide a helpful taxonomy of stakeholders in the *ML ecosystem*, distinguishing between creators, data-subjects, operators, executors, decision-subjects, and examiners.
2 Although Humphreys (2009) and several others claim that some ML models are *essentially* opaque, the present discussion is agnostic with respect to this claim. That is, it only concerns models that can in fact be rendered transparent through Explainable AI, however numerous these may be.

The following discussion describes three ways in which Explainable AI facilitates scientific exploration. Section 2 shows that some XAI techniques are well-suited for determining what ML models are models *of,* and thus, for assessing a model's suitability for a particular target. Section 3 shows that other XAI techniques can be used for causal inference, and thus, for specifying starting points for future inquiry into the causes of a particular event. Finally, section 4 shows how Explainable AI can be used to generate novel hypotheses about the algorithms that are implemented in biological brains, and thus, to provide potential explanations.

Importantly, in each one of these ways, XAI techniques' exploratory contributions can be distinguished from the contributions of the ML models to which these techniques are applied. Thus, more than just being a solution to the problem that opacity poses, Explainable AI enhances the overall exploratory potential of machine learning and data-driven scientific inquiry.

2. Determining What a Model is a Model *Of*

In a recent commentary, Emily Sullivan (2019) examines the use of ML models in scientific research. Although she denies that opacity negatively impacts these models' scientific utility, Sullivan argues that their *link uncertainty* does. Sullivan defines link uncertainty as "a lack of scientific and empirical evidence supporting the link that connects the model to the target phenomenon" (Sullivan 2019: 1). In other words, link uncertainty arises when it is unclear what a model is a model *of*. As an illustrative example, Sullivan considers *Deep Patient*: a DNN that learns to map patients' features onto likely diseases (Miotto et al. 2016). Her point is to argue that, although the network issues reliable diagnostic predictions, the understanding that medical scientists can acquire from this model is limited. This is because it is unclear whether the model tracks genuinely causal relationships between patient features and likely diseases, or whether it is merely exploiting spurious correlations grounded in (for example) the fact that patients with certain features are tested more frequently than others.

Although Sullivan distinguishes link uncertainty from opacity, it is more appropriate to consider link uncertainty a special kind of opacity. Recall that a model is opaque when it is unclear why the model does what it does or how it works. Sullivan's discussion only concerns a lack of knowledge about how a model works. In particular, it is concerned with a lack of knowledge about a model's implementation in some particular programming language—an epistemic state that is all but guaranteed by the software-engineering practice of *encapsulation* (Mitchell 2002). This "implementation opacity" is problematic for expert creators (e.g., software developers) tasked with intervening on a model to improve its performance or to fix a bug. However, it is unproblematic for non-expert decision-subjects (e.g., medical patients) and examiners (e.g., governmental regulators), neither of which would know what to do with knowledge of a model's implementation even if they had it.

That said, stakeholders such as decision-subjects and examiners are also affected by opacity, albeit one that centers on questions about why a model does what it does, rather than on questions about how it works. Questions of this kind are answered not by specifying details of the model's implementation, but by

justifying the model's behavior through *reasons* (Zerilli et al. 2018). Unlike a model's implementation details, which concern the syntactic structures specified in a computer program, reasons in this context are individuated semantically, by reference to the environmental features and regularities that the model has learned to track (Zednik 2019). Thus, the reason why Deep Patient predicts that type-2 diabetes is likely to develop in a particular patient may be that the patient is overweight (a good reason), or that she is of advanced age (a bad reason). When Sullivan writes about link uncertainty, she is referring to a particular kind of opacity: an inability to understand the reasons for an ML model's predictions.

Given this analysis, Sullivan's claim that "implementation opacity" does not negatively impact scientific research is unsurprising: Scientific investigators are more like examiners than creators. They do not generally require knowledge of how a model works. Rather, they are interested in understanding why it does what it does. For this reason, although a lack of implementation knowledge is no obstacle to scientific research, link uncertainty is.

But of course, exposing link uncertainty as a special kind of opacity is little more than a verbal clarification. Far more important is the question of whether (and if so how) this particular kind of opacity might eventually be overcome. Can Explainable AI help scientific investigators determine what a model is a model *of*? Moreover, to what extent does overcoming this kind of opacity contribute to scientific exploration?

Many XAI techniques specialize in providing semantically-individuated reasons for a particular model's outputs. Most notably, these include techniques for identifying the input elements—be they pixels in an image or values in a table—that bear a high responsibility for a particular output. For example, visualization techniques such as *Prediction Difference Analysis* (PDA, Zintgraf et al. 2017) allow investigators to understand the regularity that is being tracked by visually inspecting a heatmap. Do the highlighted pixel regions for a model of cancerous melanoma generally look like the features that are actually characteristic of cancerous melanoma, or do they look more like irrelevant (but nevertheless correlated) features such as freckles? Moreover, do the highlighted pixel regions of the model look like features that are already known to be indicators of cancerous melanoma, or do they depict hitherto unknown (but causally relevant) indicators?

Analogous non-visual techniques may be required for models trained over tabular data. For example, *Shapley Additive Explanation* (SHAP, Lundberg & Lee 2017) ranks a model's input variables by their relative importance for producing specific outputs. Do Deep Patient's predictions of type-2 diabetes depend more on (causally relevant) factors such as a patient's weight and family background, or on (spuriously correlated) factors such as age? Moreover, do the predictions depend on factors whose relevance for type-2 diabetes is already known, or do they depend on factors whose relevance has thus far gone unrecognized? Notably, because the model's input elements (e.g., pixel regions and table values) correspond to features of its environment (e.g., skin discoloration and patient features), they can be viewed as semantically-individuated reasons for the model's outputs. Insofar as techniques such as PDA and SHAP let investigators understand these reasons, they allow them to understand what an ML model is a model *of*. In

this sense, these techniques can be used to combat link uncertainty.

Notably, Sullivan herself mentions some of these techniques in passing. Nevertheless, she stops short of recognizing their full significance for scientific exploration. In particular, although Sullivan argues that heatmaps are useful for "determining the suitability of the model" (Sullivan 2019: 25) because they can allow investigators to determine which regularity it has learned to track, she does not recognize that these techniques can also be used to identify such regularities in the first place. Indeed, ML models are renowned for their ability to uncover subtle and unintuitive regularities that would be difficult to uncover otherwise. By using techniques such as PDA and SHAP to better understand what ML models are models *of*—that is, to identify the regularities they have learned to track—scientific investigators can discover previously unknown regularities in the environment.

3. Enabling Causal Inference

The examples of link uncertainty mentioned by Sullivan are ones in which it is unclear whether the model has learned to track causal relationships as opposed to spurious correlations. But although XAI techniques such as PDA and SHAP allow investigators to determine which particular regularity is being tracked, they do not help determine whether any particular regularity is in fact a causal regularity. Put differently, these techniques do not enable causal inference.

Other XAI techniques can be used for exactly this purpose. Consider techniques that provide what Wachter et al. (2018) call *counterfactual explanations*. Counterfactual explanations specify possible worlds in which variations in a model's input yield non-actual (and possibly, desirable) outputs. A recent software tool for providing counterfactual explanations is the *Counterfactory*.[3] Given a model and input, this tool generates counterfactuals of arbitrary closeness (distance to actual input values) and complexity (number of input variables) to produce a desired but non-actual output. Thus for example, given a bank's credit-scoring model, the Counterfactory might generate counterfactuals for achieving an improved credit score: increasing income, decreasing monthly expenses, or some combination of both.

XAI techniques for counterfactual explanation can be used for causal inference, that is, for inferring the cause(s) of a particular effect. To understand how, it is worth briefly reviewing the close connection between counterfactual reasoning and causal inference. Consider an actual scenario in which event *C* (e.g. the striking of a match) precedes event *E* (e.g. the match catching fire), over an arbitrary number of background conditions *B* (e.g. the surrounding temperature being 19°C, there being oxygen in the air, etc.). Assuming that all *B* remain constant, one can infer that *C* is causally relevant for *E* if and only if a counterfactual change in *C* co-occurs with a change in *E*.

Causal inference can serve the purposes of many different stakeholders. Decision-subjects can assume a degree of control over model-driven decisions if they can

3 Proprietary technology currently being developed by the neurocat GmbH: https://www.neurocat.ai/ (retrieved August 18th, 2020).

infer the changes to make so as to effect a different model output (e.g., whether they need to earn more to improve their credit score). Examiners can assess a model's compliance with ethical or legal norms if they can determine the causal relevance of certain key variables (e.g., whether credit scoring causally depends on gender or ethnicity). More relevant in the present context, scientific investigators can engage in causal inference to determine whether the regularity being tracked by a model is in fact a causal regularity. If a software tool can generate counterfactuals in which a change in *E* is predicted from a change in *C,* investigators might infer (assuming all *B* remain equal) that the learned relationship between *C* and *E* is genuinely causal as opposed to merely correlative.

Of course, the differences between the industrial and scientific contexts are significant. In industry, what matters is (typically) the model itself. In such contexts, XAI techniques for counterfactual explanation are perfect guides to causal inference: If the Counterfactory generates a counterfactual in which a higher income yields an improved credit score, then a higher income will actually yield an improved credit score. In science, by contrast, what matters is (typically) the domain that the model is a model *of*. Accordingly, in these contexts, XAI techniques for counterfactual explanation are imperfect guides to causal inference: If the Counterfactory generates a counterfactual in which losing weight yields a reduced probability of type 2-diabetes, then it is still possible that losing weight does not actually reduce the probability of type-2 diabetes. Because scientific models can be false, the causal inferences grounded on these models are insecure.

That said, the insecurity of XAI-driven causal inference does not render it useless for scientific research. On the contrary, it can serve an invaluable exploratory purpose. In particular, XAI techniques for counterfactual explanation can be used to refine extant causal hypotheses as well as to generate new ones. Consider the hypothesis that excessive weight is causally relevant for type 2-diabetes. This is a well-confirmed hypothesis, despite the fact that many overweight people never actually become diabetic (Wu et al. 2014). Nevertheless, it may be desirable to subsume the exceptions under a more-refined hypothesis. Indeed, applying the Counterfactory to Deep Patient might suggest suitable refinements. For example, counterfactuals generated for a desired outcome of less-probable diabetes might combine weight-loss with an additional factor, such as an absence of sleep apnea. Motivated by these counterfactuals, scientists might conduct further experiments, and if necessary, refine the original hypothesis so that excessive weight is only deemed causally relevant when it co-occurs with sleep apnea. In this (admittedly hypothetical) scenario, XAI-driven causal inference identified a starting point for scientific inquiry: generating new hypotheses, devising potential explanations, and inspiring new experiments.

Notably, XAI-driven causal inferences can perform this exploratory function in almost any scientific domain in which ML models have been developed for predictive purposes. In synthetic biology, for example, investigators may deploy such inferences to identify and test genetic modifications that are likely to yield desirable phenotypic traits (Ma et al. 2018). Analogously, in chemistry they might use XAI techniques for counterfactual explanation to discover new compounds with desirable (e.g., pharmaceutical) properties (Zhavoronkov 2018). Given the

increasingly important role that machine learning plays in many different scientific domains, the exploratory promise of XAI-driven causal inference is tantalizing.

Before moving on, it is worth dwelling briefly on the kinds of domains for which XAI-driven causal inference might be particularly useful. Software tools such as the Counterfactory are remarkably efficient even for high-dimensional nonlinear DNNs, can be applied to any model-type and a wide variety of use-cases, and can generate counterfactuals even for intrinsically high-dimensional data-types such as naturalistic images. Given that ML models are capable of tracking high-dimensional and nonlinear regularities in complex systems such as the brain or the climate, such tools (assuming the relevant model is approximately true) might facilitate causal inference even for systems of such high levels of complexity. If true, this would be a significant achievement indeed: high-dimensionality and nonlinearity are among the biggest obstacles for traditional causal inference methods, which tend to work well only when the variables are few and the relationships are linear (Bühlmann 2013). Insofar as ML models can be trained to replicate the behavior of ever larger and more complex systems, and insofar as XAI techniques can be used to counterfactually explain the behavior of these models, Explainable AI is poised to significantly extend the limits of causal inference.

4. Generating Algorithmic-Level Hypotheses

Techniques from Explainable AI can perform at least one more exploratory role: generating *algorithmic-level hypotheses* that serve as potential explanations. The notion of an algorithmic-level hypothesis requires elaboration. Some physical systems—most notably biological brains—are computational systems insofar as they perform computational tasks in their surrounding environments (Shagrir 2006). Although these systems can be described at a physical level of analysis, by specifying the spatiotemporal structures and processes that underlie their behavior, it is often more insightful to describe them at an *algorithmic* level of analysis, by specifying the algorithms they execute in the service of the task (Marr 1982). Indeed, cognitive science is to a large extent in the business of formulating testable hypotheses about the structure, efficiency, and representational content of algorithms that biological organisms use to accomplish cognitive tasks such as perception, categorization, memory-formation, and language-learning. Notably, although many such hypotheses have been articulated and evaluated in the past, there is no general agreement about the way in which new algorithmic-level hypotheses should be developed in the future. To a certain extent, cognitive modeling remains an inscrutable “dark art”.

Explainable AI may help transform this “dark art” into a semi-autonomous exploratory process. Specifically, XAI techniques can facilitate the specification of algorithms to test as possible explanatory hypotheses. Indeed, given that many ML models are trained to perform tasks that closely resemble the ones that are performed by biological cognizers, and given that these models are often trained on naturalistic datasets that mirror the real-world environments in which those cognizers develop and learn, it is at least not wholly unreasonable to assume that ML models might implement algorithms that bear at least some similarity to the algorithms that are implemented in biological brains (see also Zednik 2018). Insofar as XAI techniques allow cognitive scientists to understand and describe the

algorithms that are learned by a particular model, they can also be used to articulate new and hitherto unconsidered hypotheses about the algorithms that are learned by biological brains.

At this point, it may be necessary to clarify why XAI should be necessary at all, within the context of understanding the algorithms that are learned by ML-programmed models. Although human programmers typically decide on a model's *learning* algorithm, they have limited influence on the structure and function of what might be called the *learned* algorithm. For example, although they might train a DNN using some variant of the backpropagation algorithm, they do not determine the values that this algorithm (when applied to a particular learning environment) eventually assigns to individual network parameters (e.g., connection weights). Since it is these parameters that govern the model's output for any particular input, they implement a learned algorithm for computing a particular function. But what exactly this algorithm is, and how it might be characterized in a concise, understandable (and potentially modifiable) way, is obscured by the fact that the number of network parameters is high and their interdependencies are nonlinear.

Notably, whereas the XAI techniques considered in previous sections serve to answer questions about why an ML model does what it does by specifying reasons, the techniques to be considered here answer questions about how such a model works by uncovering algorithms. One way of uncovering algorithms is by using any one of a diverse family of *surrogate modeling* techniques. These techniques specify (relatively) simple algorithms to replicate (to an arbitrary degree of precision) an opaque model's overt behavior and internal processing. In particular, *rule-extraction* methods (e.g., Zilke et al. 2016) produce rule lists that approximate the input-output behavior of any high-dimensional DNN. Similarly *tree-extraction* methods (e.g., Wu et al. 2018) produce decision-trees that replicate the internal decision-structure of complex and (even recurrent) neural networks.

Intriguingly, these surrogate models bear a structural resemblance to classic "symbolic" models that were used widely in cognitive science throughout the 1960s, 70s and 80s. Because some of these models remain in use today, it is not unreasonable to suppose that surrogate models for explaining the behavior of trained ML models might be advanced as candidate hypotheses for explaining the behavior of biological cognizers. That said, many areas of cognitive science have by now moved on to "subsymbolic" methods that more closely resemble the methods commonly used by neuroscientists. Indeed, some of these methods may even serve double-duty, simultaneously explaining the behavior of biological brains and of artificial neural networks.

Consider, for example, *representational similarity analysis* (RSA, Kriegeskorte & Kievit 2013; Kriegeskorte et al. 2008). RSA is an integrative technique for data-analysis that lets neuroscientists relate multi-channel brain-activity data to each other, to behavioral data, to data produced by conceptual and computational models, and to stimulus descriptions by comparing (representational) dissimilarity matrices (RDMs). Cichy et al. (2016) have recently deployed this technique to compare temporal and spatial brain representations with representations in a deep feed-forward neural network trained for object categorization. That is, they

aim to use RSA to identify a DNN's learned representations for object-recognition, and to determine whether these representations bear a structural similarity to the brain's representations in an analogous task.

How exactly is this aim achieved? First, for each signal space (DNN, fMRI, and MEG) Cichy et al. estimate the representational activity patterns associated with 118 experimental stimuli (images of natural objects over real-world backgrounds). Second, for each signal space of every pair of experimental stimuli, they compute the activity pattern dissimilarity. This yields 118-by-118 RDMs (each one of which contains the dissimilarity values for all experimental stimuli-pairs) for every DNN layer, every fMRI region-of-interest or searchlight, and every millisecond in the MEG signal. Third, DNN RDMs are directly compared to fMRI or MEG RDMs by calculating the Spearman rank correlation coefficients between them, yielding a relatively easy measure of brain-DNN representational similarity. In this way, RSA permits a specification of the representations that are used by both the DNN and the brain, and a subsequent comparison of these representations at the level of RDMs.

Indeed, the comparison reveals that "the DNN captured the stages of human visual processing in both time and space from early visual areas towards the dorsal and ventral streams" (ibid.: 1). Moreover, a close analysis of the representational structures in the DNN supports a series of specific empirical predictions:

> "Our results demonstrate the explanatory and discovery power of the brain-DNN comparison approach to understand the spatio-temporal neural dynamics underlying object recognition. They provide novel evidence for a role of parietal cortex in visual object categorization, and give rise to the idea that the organization of the visual cortex may be influenced by processing constraints imposed by visual categorization the same way that DNN representations were influenced by object categorization tasks." (ibid.: 9)

Overall, although (or perhaps because) RSA was originally developed by neuroscientists to investigate representations in the brain, this technique may not only be used to explain the behavior of trained neural networks, but also to generate and test algorithmic-level hypotheses about biological brains. Notably, in this particular case, the generated hypothesis seems likely to be confirmed, suggesting that XAI may not only facilitate exploration, but also explanation.

5. Conclusion

Models developed using Machine Learning are assuming an increasingly prominent place in scientific research. Many recent discussions recognize the problem that opacity poses to the use of such models, and some of these discussions have begun to reflect on the possibility of solving this problem through the use of Explainable AI. However, Explainable AI appears to be more than just a solution to a problem. This paper has sought to show that XAI techniques can serve an invaluable exploratory role in their own right, over and above the ML models to which these techniques are applied.

In particular, tools such as PDA and SHAP have been shown to answer questions about why a model does what it does. Thus, they allow scientific investigators to better understand what a model is a model *of*, and to assess its suitability for a particular target. Moreover, XAI techniques for counterfactual explanation have been shown to enable causal inference—perhaps even over domains that are at once high-dimensional and nonlinear. In this way, these techniques reveal new starting points for scientific inquiry: new hypotheses to test, and new experiments to conduct. Finally, surrogate modeling techniques and analytic techniques such as RSA can be used to better understand the algorithms and representations that are learned by models to accomplish particular tasks. Insofar as there is reason to believe that these algorithms might also be implemented in biological brains, they can be advanced as potential explanations in cognitive science. For all of these reasons and more, Explainable AI is a promising new tool for scientific exploration.

6. References

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251.

Bühlmann, P. (2013). Causal statistical inference in high dimensions. *Mathematical Methods in Operations Research, 77*(3), 357–370.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

Cichy, R. M. & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23(4), 305–317.*

Dattilo, A. et al. (2019). Identifying exoplanets with deep learning II: Two new super-earths uncovered by a neural network in K2 data. *arXiv*, 1903.10507.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv*, 1710.00794.

Gelfert, A. (2016). *How to do science with models. A philosophical primer.* Springer: Dordrecht.

Hohman, F. M., Kahng, M., Pienta, R., & Chau, D. H. (2018). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*(4), 1–28.

Kriegeskorte, N. & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401–412.

Lundberg, S. M. & Lee, S. (2017). A unified approach to interpreting model predictions. *arXiv*, 1705.07874v2.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., & Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta, 248*(5), 1307–1318.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Miotto, R., Li, L., Kidd, B. A. and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*(1), 1–10.

Mitchell, J. C. (2002). *Concepts in programming languages*. Cambridge: Cambridge University Press.

Ritchie, J. B., Kaplan, D.M. & Klein, C. (2019). Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science* 70(2): 581-607.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Schmidt, J., Marques, M. R. G., Botti, S. *et al.* (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials, 5*, 83.

Shagrir, O. (2006). Why we view the brain as a computer. *Synthese*, *153*(3): 393–416.

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, axz035.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv*, 1806.07552.

Wachter, S., Mittelstadt, B. & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2).

Wu, Y., Ding, Y., Tanaka, Y. & Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *International Journal of Medical Sciences* 11(11): 1185-1200.

Wu, M., Hughes, M. C., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. *arXiv*, 1711.06178v1.

Zednik, C. (2018). Will machine learning yield machine intelligence? In V. Müller (ed.) *Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017. Studies in*

Applied Philosophy, Epistemology and Rational Ethics, 44. Springer: Cham

Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, *32*(4), 661–683.

Zhavoronkov, A. (2018). Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. *Molecular Pharmaceutics,15*(10), 4311-4313.

Zilke, J. R., Mencia, E. L., & Janssen, F. (2016). DeepRED – Rule extraction from deep neural networks. In T. Calders, M. Ceci, D. Malerba (Eds.): *Discovery Science 19th International Conference* (pp. 457–473).

Zintgraf, L. M., Cohen, T. S., Adel, T,. & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *The fifth International Conference on Learning Representations* (ICLR).